The interplay of descriptor-based computational analysis with pharmacophore modeling builds the basis for a novel classification scheme for feruloyl esterases

D.B.R.K. Gupta Udatha[1,2], Irene Kouskoumvekaki[2], Lisbeth Olsson[1] & Gianni Panagiotou[1,2]

One of the most intriguing groups of enzymes, the feruloyl esterases (FAEs), is ubiquitous in both simple and complex organisms. FAEs have gained importance in biofuel, medicine and food industries due to their capability of acting on a large range of substrates for cleaving ester bonds and synthesizing high-added value molecules through esterification and transesterification reactions. During the past two decades extensive studies have been carried out on the production and partial characterization of FAEs from fungi, while much less is known about FAEs of bacterial or plant origin. Initial classification studies on FAEs were restricted on sequence similarity and substrate specificity on just four model substrates and considered only a handful of FAEs belonging to the fungal kingdom. This study centers on the descriptor-based classification and structural analysis of experimentally verified and putative FAEs; nevertheless, the framework presented here is applicable to every poorly characterized enzyme family. 365 FAE-related sequences of fungal, bacterial and plantae origin were collected and they were clustered using Self Organizing Maps followed by k-means clustering into distinct groups based on amino acid composition and physico-chemical composition descriptors derived from the respective amino acid sequence. A Support Vector Machine model was subsequently constructed for the classification of new FAEs into the pre-assigned clusters. The model successfully recognized 98.2% of the training sequences and all the sequences of the blind test. The underlying functionality of the 12 proposed FAE families was validated against a combination of prediction tools and published experimental data. Another important aspect of the present work involves the development of pharmacophore models for the new FAE families, for which sufficient information on known substrates existed. Knowing the pharmacophoric features of a small molecule that are essential for binding to the members of a certain family opens a window of opportunities for tailored applications of FAEs.

The carbohydrates of lignocelluloses and lignin are covalently linked by plant aromatics or phenolics and at times physically mask the potentially fermentable substrates from degradation and bioconversion (Akin, 2008). The presence of phenolic acid esterases such as ferulic acid esterases or feruloyl esterases (FAEs) (E.C. 3.1.1.73) enables microorganisms to attack and partially degrade aromatic-containing plant tissues. Also the rumen ecosystem, with diverse microorganisms that produce cocktails of enzymes, successfully degrades plant biomass which is the source of ruminant animals' energy. The usefulness of FAEs in pulp and paper sector industries has been also reported with their potentiality of cleaving the covalent links between hemicellulose and pectin to aromatic compounds of lignin (Record et al, 2003). FAEs have gained importance in biofuel industry due to their capability of enhancing the accessibility of plant tissues to cellulolytic and hemicellulolytic enzymes. With the importance of FAEs along with other cellulolytic enzymes in biomass degradation, construction of bifunctional enzymes as improved enzymatic tools to degrade agricultural by-products has also been demonstrated (Levasseur et al, 2005). Ferulic acid, one of the most abundant hydroxycinnamic acid liberated from the action of FAEs on agricultural by-products, has gained importance in food industry as it can be further transformed from a variety of microorganisms into vanillin, a flavouring food additive (Lesage-Meessen et al, 1996). Different other types of hydroxycinnamic acids liberated from FAEs have importance in cosmetic and pharmaceutical industries due to their antioxidant properties (Kikuzaki et al, 2002).

During the last decade, FAEs have gained increased attention in the area of biocatalytic transformations for the synthesis of hydroxycinnamic acid esters with medicinal and nutritional applications. Feruloylation of D-arabinose by a FAE and its potential application as anti-mycobacterial agent has recently been demonstrated (Vafiadi et al, 2007b). Furthermore, the potential of

[1]Department of Chemical and Biological Engineering, Industrial Biotechnology, Chalmers University of Technology, Kemivägen 10, SE-412 96, Gothenburg, Sweden. [2]Center for Biological Sequence Analysis, DTU-Systems Biology, Building 208, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark. Correspondence should be addressed to G.P. (gpa@bio.dtu.dk).

FAE as a synthetic tool of various phenolic esters and their inhibitory effect on LDL (Low-Density-Lipoproteins) oxidation has been investigated *in vitro* towards the prevention of atherosclerosis (Vafiadi et al, 2008). Studies on bioavailability of diferulic acids based on their intestinal release and uptake of phenolic antioxidants have been carried out showing that esterified diferulates can be released from cereal bran by intestinal enzymes (microflora) and the free diferulic acids can enter the circulatory system (Andreasen et al, 2001). Epidemiological and laboratory studies have also shown that dietary administration of cereal bran (Slavin, 2000) protect against colon tumorigenesis (Ferguson et al, 1999).

Having a wide range of demonstrated applications, the amount of research on FAEs has been increased in the last decade and FAEs from various microorganisms with different reaction specificities have been published. Several studies have been reported on the isolation, purification and partial characterization of FAEs from fungi (Castanares and Wood, 1992; Donaghy and McKay, 1997; Koseki et al, 1998; Garcia-Conesa et al, 2004). Following the increasing attention on these enzymes, fungal FAEs were classified into four different types (Table 1) based on sequence homology and substrate specificity of seven enzymes (Crepin et al, 2004). However, a multiple sequence alignment analysis of the fungal FAEs by Crepin et al (2004) led to the assumption of a putative type 'E' with unknown biochemical characteristics. Making a step further, seven subfamilies of FAEs were proposed based on phylogenetic analysis of fungal FAEs by Benoit et al (2008), but the subfamilies included characterized FAEs from Type A, B and C only; Type D FAEs were absent in the above analysis. The difference between subfamilies in the above classification system was phylogenetic rather than functional, where possibly gene loss and gene duplication events were behind the annotation of a FAE to one of the seven subfamilies (Benoit et al, 2008).

Table 1 Classification of fungal feruloyl esterases by Crepin et al (2004) based on specificity for methyl esters and sequence homology.

Type	Substrate specificity	Ability of releasing free diferulates	Sequence homology
A	Ability to hydrolyse Methyl Ferulate, Methyl Sinapate and Methyl p-coumarate	5,5' - diferulic acid only	Lipase
B	Ability to hydrolyse Methyl Ferulate, Methyl Sinapate and Methyl Caffeate	No	Cinnamoyl esterase family 1 and Acetyl xylan esterase
C	Ability to hydrolyse Methyl Ferulate, Methyl Sinapate, Methyl p-coumarate and Methyl Caffeate	No	Chlorogenate esterase and Tannase
D	Ability to hydrolyse Methyl Ferulate, Methyl Sinapate, Methyl p-coumarate and Methyl Caffeate	5,5' - diferulic acid only	Xylanase

In addition, data regarding the hydrolytic and synthetic specificity of FAEs have been generated via a variety of methods and procedures. For assaying the FAE activity, researchers have used different model substrates (Giuliani et al, 2001; Topakas et al, 2003a; Topakas et al, 2003c; Hatzakis et al, 2003; Topakas et al, 2004; Topakas et al, 2005a; Vafiadi et al, 2005; Vafiadi et al, 2006; Tsuchiyama et al, 2006; Vafiadi et al, 2007a; Vafiadi et al, 2007b; Tsuchiyama et al, 2007; Vafiadi et al, 2008; Vafiadi et al, 2009, Goldstone et al, 2010) and the information available from recent works on hydrolytic (Supplementary File 1 - Table S1) and synthetic specificity (Supplementary File 1 - Table S2) of FAEs have challenged the previously proposed classification system that was based on the specificity for only four substrates (Crepin et al, 2004). For example, StFaeB, a FAE from *Sporotrichum thermophile* and FoFaeB, a FAE from *Fusarium oxysporum* fall under type-B of Crepin's FAE classification, however, the two FAEs showed different substrate specificities against 26 methyl phenylalkanoates (Topakas et al, 2004). Similarly, TsFaeC from *Talaromyces stipitatus* and StFaeC from *Sporotrichum thermophile* that were classified before as type-C have also shown different specificities (Vafiadi et al, 2006).

Up to now only a handful of fungal FAEs have been classified based on their specificity for four aromatic substrates, their varying ability to release diferulic acids from esterified substrates (Crepin et al, 2004) and their sequence homology (Benoit et al, 2008). Reports were also published on FAEs from bacteria with different specificities (Donaghy et al, 2000; Blum et al, 2000; Bartolomé et al, 2003; Wang et al, 2004; Wang et al, 2005; Laszlo et al, 2006; Mukherjee et al, 2007; Aurilia et al, 2007; Rashamuse et al, 2007; Nsereko et al, 2008; Kheder et al, 2009; Dodd et al, 2009), while a few reports exist on feruloyl esterases from plants (Sancho et al, 1999; Humberstone and Briggs, 2002; Tomoko et al, 2002; Latha et al, 2007). To the best of our knowledge, there is no classification system that has been reported in the literature with a common platform considering FAEs and related enzymes from all the three kingdoms. Few enzymes like lipases, cutinases and tannases have shown similar activities to the FAEs (García-Conesa et al, 2001; Andersen et al, 2002), which suggests that FAE related enzymes should also be considered in the classification schemes.

We observed that the available sequenced and characterized FAEs belong to different protein superfamilies covering lipases, tannases and feruloyl esterases. The list of these characterized FAEs and their respective superfamilies, is shown in Table 2. These recent works have made apparent the need for a new classification system for FAE.

Table 2 Superfamilies of characterized FAE sequences

GI number & Protein accession	Species	Kingdom	Enzyme	Superfamiliy [a]
gi|17366177|sp|O42807.1	*Aspergillus niger*	Fungi	Feruloyl esterase (FaeA)	Lipase superfamily
gi|17932783|emb|CAC83933.1|	*Aspergillus niger*	Fungi	Feruloyl esterase (FaeB)	Tannase superfamily
gi|9955721|emb|CAC05587.1|	*Neurospora crassa*	Fungi	Feruloyl esterase (Fae-I)	Esterase_lipase superfamily
gi|17366179|sp|O42815.1|	*Aspergillus tubingenesis*	Fungi	Feruloyl esterase (FaeA)	Lipase superfamily
gi|33945411|emb|CAD44531.1|	*Talaromyces stipitatus*	Fungi	Feruloyl esterase (TsFaeC)	Tannase superfamily
gi|84028205|sp|Q9P979.2|	*Aspergillus awamori*	Fungi	Feruloyl esterase (AwFaeA)	Lipase superfamily
gi|7839348|gb|AAF70241.1|	*Orpinomyces sp. PC-2*	Fungi	Feruloyl esterase (FaeA)	Esterase_lipase superfamily
gi|23821548|sp|Q9Y871.1|	*Piromyces equi*	Fungi	Feruloyl esterase (EstA)	CBM_10 & Esterase_lipase superfamily
gi|30315043|gb|AAP30751.1|	*Piromyces sp E2*	Fungi	Feruloyl esterase (FaeA)	CBM_10 superfamily
gi|25090320|sp|Q9HE18.1|	*Penicillium funiculosum*	Fungi	Feruloyl esterase (FAE-I)	Esterase_lipase superfamily
gi|67522631|ref|XP_659376.1|	*Aspergillus nidulans*	Fungi	Feruloyl esterase (FaeB)	Tannase superfamily
gi|83766486|dbj|BAE56626.1|	*Aspergillus oryzae*	Fungi	Feruloyl esterase (AoFaeB)	Tannase superfamily
gi|83766949|dbj|BAE57089.1|	*Aspergillus oryzae*	Fungi	Feruloyl esterase (AoFaeC)	Tannase superfamily
gi|32420917|ref|XP_330902.1|	*Neurospora crassa*	Fungi	Feruloyl esterase (NcFaeD)	Esterase_lipase superfamily
gi|74271753|dbj|BAE44304.1|	*Penicillium chrysogenum*	Fungi	Feruloyl esterase (FAE-I)	Tannase superfamily
gi|18159028|pdb|1GKK|A	*Clostridium thermocellum*	Bacteria	Feruloyl Esterase Domain Of Xyny	Esterase_lipase superfamily
gi|16974939|pdb|1JJF|A	*Clostridium thermocellum*	Bacteria	Feruloyl Esterase Domain Of XynZ	Esterase_lipase superfamily

[a] Superfamily designations are according to NCBI Conserved Domains Database (CDD). They are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences using "Reverse Position-Specific BLAST". CDD consists of NCBI-curated domains and domain models from several external source databases (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml)

In this research review paper we applied an array of computational tools and we succeeded to develop a new classification scheme for FAEs, which opens new vistas in the application of this intriguing group of enzymes. The present work is not restricted to FAEs but represents a framework for the functional characterization and identification of substrate specificity for any poorly characterized enzyme group. In addition we demonstrate that sequence information can be used to develop models that are able to elucidate the underlying structural characteristics that determine substrate specificities.

Phylogenetic analysis of FAEs

Every type of sequence analysis method (e.g. evolutionary analysis, protein secondary structure predictions, etc) relies on Multiple Sequence Alignment (MSA) for similarity searches and phylogenetic analysis. Initially, we conducted phylogenetic analysis of FAEs based on multiple sequence alignment of FAE sequences. Multiple sequence alignment aims to draw a bird's eye view for common evolutionary origin in the same column position for the given set of sequences. Usually sequence aligners depend on classic methods such as ClustalW (Thompson et al, 1994) that generate fast and reproducible results, which made it one of the most widely cited paper (>30,000 citations) in biology. In the present study, we used ClustalW for multiple sequence alignments and the latest version (version 2.0) of ClustalX (Larkin et al, 2007) for bootstrapping phylogenetic trees. Sequence alignments unambiguously distinguish between protein pairs of similar and non-similar structure when the pair-wise sequence identity is greater than 40% for long alignments (Rost, 1999). The signal gets blurred in the twilight zone of 20-35% sequence identity and the alignments become messy with long gaps. Alignments after removing signal peptides may provide better conclusions on evolutionary analysis (Wilkinson et al, 2005). The presence of signal peptides complicates the prediction of amino acid sequence properties, thus, if present, it is recommended to remove those (Lao et al, 2002). We removed the signal peptides of

each sequence after first predicting them using the SignalP 3.0 server and the respective options provided by the server depending on the eukaryotic or bacterial origin of the protein (Nielsen et al, 1997, Bendtsen et al, 2004).

An issue that we had to deal with during the early stages of data collection was that of ambiguities in protein nomenclature, where the same sequence appeared with different protein identifiers in different reference sources. For example, the FAE from *Aspergillus awamori* was reported with the DB number BAA92937 in Koseki et al, 2009b, whereas the same FAE was identified as Q9P979 in Benoit et al, 2008 and as AB032760 in Topakas et al, 2007. To make a common platform for the identification and classification of FAEs, we have used the GI sequence identification number used by NCBI (National Center for Biotechnology Information). The amino acid sequences were retrieved from the NCBI protein database using 'feruloyl esterase' as key word and the amino acid sequences that appeared more than once were filtered out. Additionally, we increased the data collection of putative FAEs by PSI-BLAST (Position-Specific Iterated BLAST) that uses position-specific scoring matrices to detect distant evolutionary sequence relationships (Altschul et al, 1997). As input for the PSI-BLAST we used the protein sequences of the seventeen experimentally characterized FAEs. MSA of FAE-related sequence hits obtained from PSI-BLAST was performed to remove identical sequences with different sequence identifiers. FAE-related sequences of *Fusarium* species were retrieved from the BROAD Institute Database (http://www.broadinstitute.org) and FAE-related sequences of *Mycosphaerella graminicola* and *Nectria haematococca* were retrieved from the DOE Joint Genome Institute Database (http://www.jgi.doe.gov/). All the sequences were further shortlisted by removal of identical sequences (using MSA-ClustalW) and sequences that are incomplete either at amino or carboxyl terminal.

The schematic diagram for retrieval and consolidation of characterized/partially characterized and FAE-related sequences is shown in Figure 1. The retrieved 365 sequences have 54% fungal, 45% bacterial and 1% plant origin. In the majority of cases more than 10 FAE-related sequences per genome were identified. However in some organisms the number of putative FAEs was either significantly higher (e.g. 27 FAE-related sequences in *Aspergillus niger*, 16 FAE-related sequences in *Nectria haematococca*) or very low (1 FAE-related sequence in *Aspergillus tubingensis*, 1 FAE-related sequence in *Piromyces equi*).

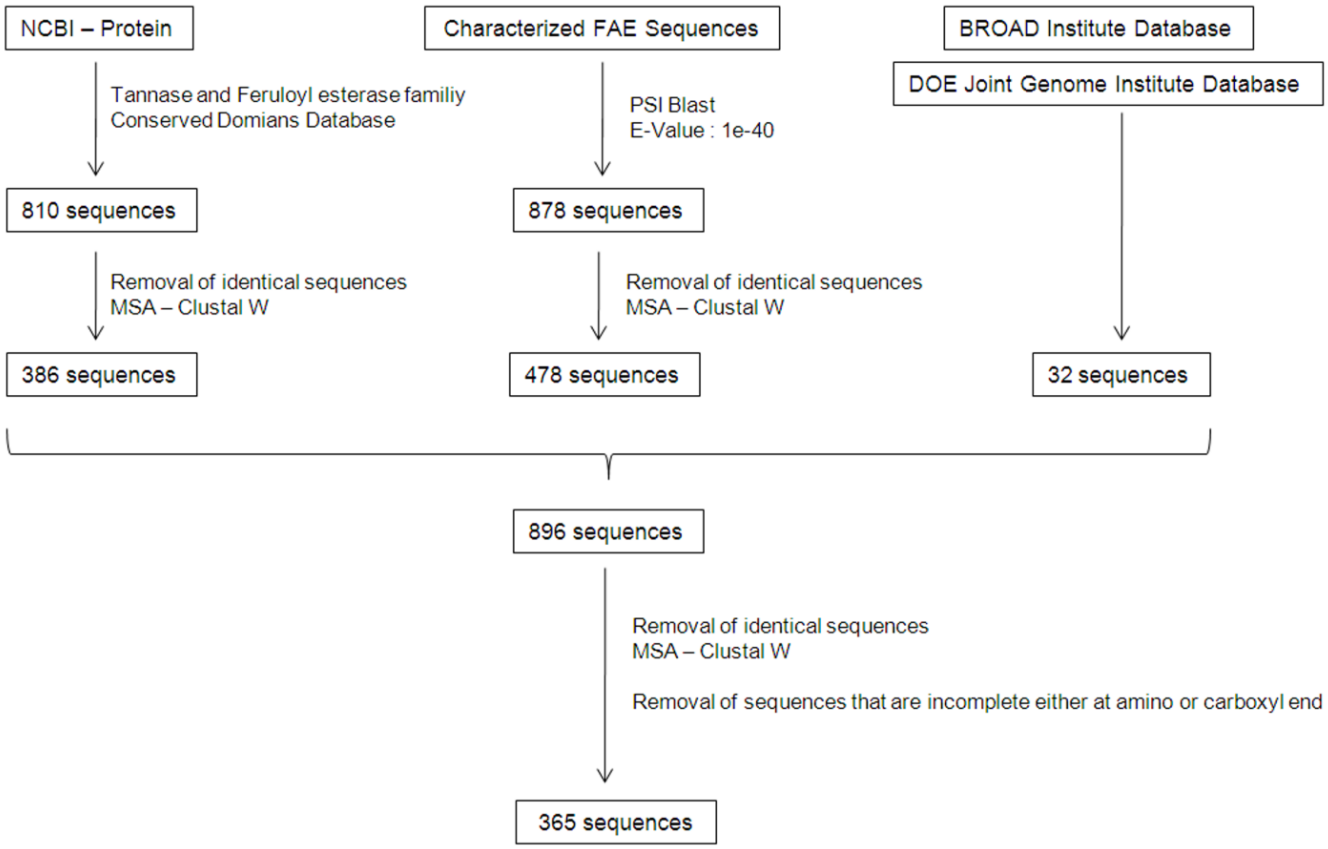


Figure 1 Scheme followed for retrieval and consolidation of characterized FAEs and FAE-related sequences. Out of 365 FAE-related sequences, 197 are from fungi, 163 are from bacteria, 4 are from plants and only one sequence is from protists.

The source organism information of each sequence was collected from NCBI (http://www.ncbi.nlm.nih.gov). All sequences with their respective GI number were further sorted according to taxonomy lineage and are listed in Supplementary File 1 (Tables S3, S4, S5 and S6). The length of signal peptides in the FAE-related sequences predicted using SignalP 3.0 server and the final length of the respective protein considered for further analysis were listed in Supplementary File 1 (Table S7).

After Multiple Sequence Alignment of the 365 FAEs and FAE-related sequences using ClustalW2, a bootstrapped Neighbor-Joining tree (N-J tree) was constructed using ClustalX Version 2.0 program. Bootstrapped N-J tree derives confidence values by making N random samples of sites from the alignment, drawing N trees (1 from each sample) and counting how many times each clade (cluster) from the original tree occurs in the sample trees. Here we used 100 bootstrap trails for constructing the tree. The distribution of FAE sequences among the clades is shown in the Supplementary File 1 (Table S8). To get an overview of this large phylogenetic tree, clades are presented in a circular phylogram (Figure 2) which was created using Dendroscope software (Huson et al, 2007). FAE-related sequences from all kingdoms are well distributed among the bootstrapped phylogram, demonstrating the close relatedness of FAEs from fungi, bacteria and plantae.

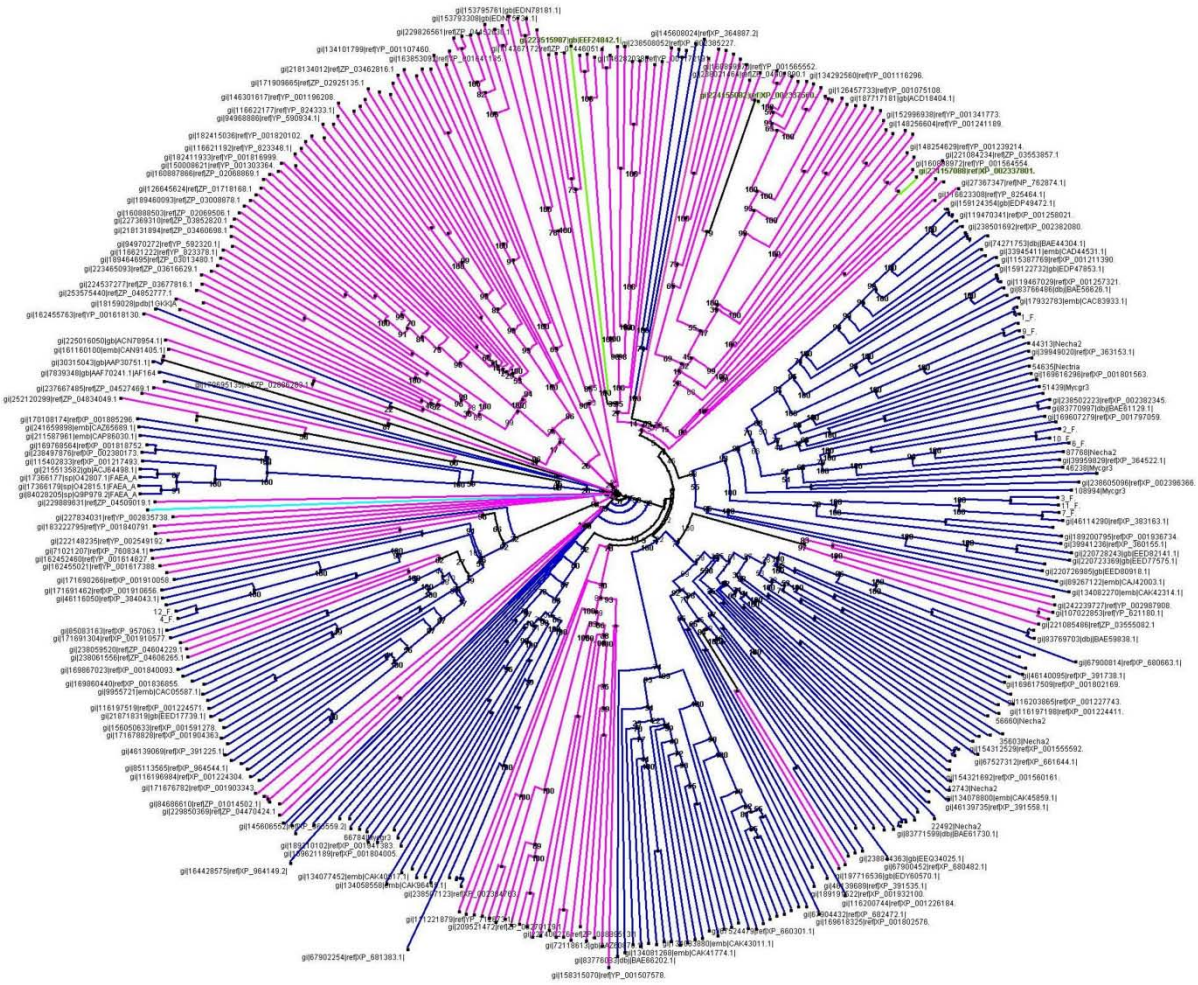


Figure 2 Circular phylogram: Bootstrapped N-J tree of 365 FAE sequences. The distributions of the FAE sequences of respective kingdoms among the clades are represented with line colours (Fungi – Blue; Bacteria – Magenta; Plants – Green; Protists – turquoise). Bootstrap confidence values are shown on the phylogram. For lineage details of the FAE sequences see Table S8 in Supplementary File 1.

The taxonomy lineage shown in Figure 2 and even more clearly at the bootstrapped rectangular phylogram of Table S8 (Supplementary File 1) is jumbled and show a patchy distribution of taxonomy lineage suggesting the probability of a lateral gene transfer, gene gain or gene loss events which might have occurred throughout the large phylogeny across fungi, bacteria and plantae. The close relationship of the plant putative FAE sequences with bacterial sequences at the tree edges of the clades in the phylogram indicates that acquisition of FAE-related genes by plantae was a relatively recent event.

Figure 3 shows the above circular phylogram with a different colour for each clade. Even though type A FAEs from *Aspergillus* species are clustered together in clade A of the phylogram, this clade also includes a type D FAE from *Neurospora crassa*. Another type D FAE from *Piromyces equi* falls under clade B. Type B FAEs are found in both clades D and J of the phylogram, while the former contains FAEs of type C as well. None of the characterized FAEs are present in clades C, E, F, G, H, I and K of the phylogram.

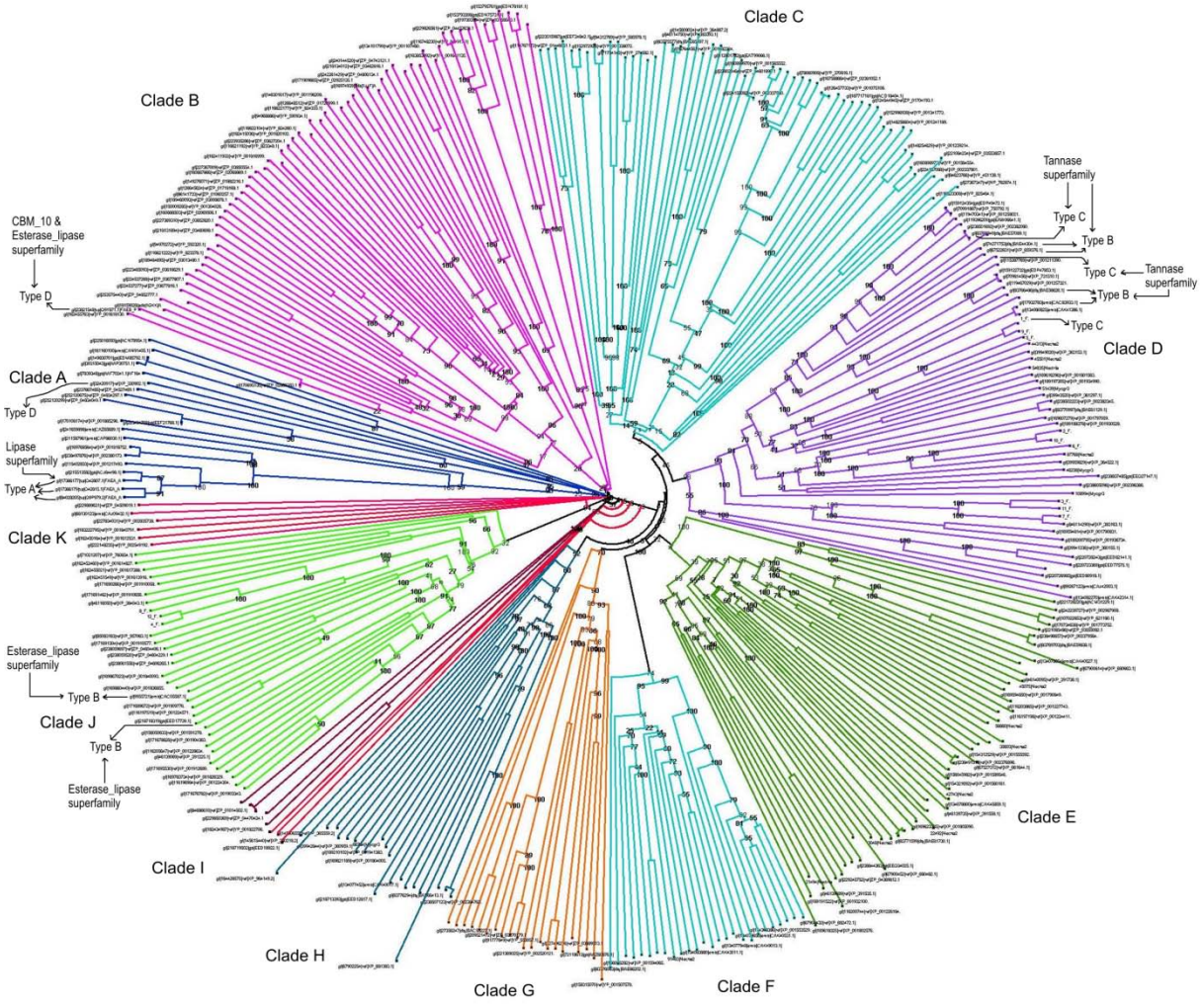


Figure 3 Circular phylogram: Bootstrapped N-J tree of putative FAE sequences. Clades are shown in different colours. Different types of characterized FAEs are spread among the clades A, B, D and J of the phylogram. FAE types are according to the classification system proposed by Crepin et al (2004). Bootstrap confidence values are shown on the phylogram.

The fact that there are FAEs of different types within the same clade as well as that FAEs of the same type are not always clustered in the same clade, suggest that phylogenetic clustering is not correlated with substrate binding specificity and, thus, it is not sufficient for the functional classification of FAEs from fungi, bacteria and plants. The above phylogenetic analysis reveals the complexity of evolutionary relationships between the FAEs of fungi, bacteria and plantae. Furthermore FAEs belong to different protein superfamilies (Table 2) with respective evolutionary histories. This observation is in line with the work of Levasseur et al (2006) that describes an unequivocal connection between evolutionary and functional shifts in fungal FAEs that were probably linked to environmental changes and might have driven adaptation by functional diversification and by molecular adaptation leading to novel enzymes. However, it should be noted that the above grouping of FAEs into different clades of the bootstrapped phylogram was based on their primary amino acid sequence identity and further conclusions cannot be drawn at this point due to the lack of biochemical data.

Classification system based on descriptors derived from sequence

Identifying the pattern of residues that cause changes in substrate specificity and developing a new classification system for sub-grouping FAEs according to function is important, as this could allow one to select the optimum FAE for a particular type of application. FAEs with the classic constellation of the Ser-His-Asp catalytic triad (McAuley et al, 2004), have evolved from a common ancestor. Different FAEs can bind to different substrates with varied degree of affinity but they often also share some common substrates. From Table 2, it is clear that FAEs arise from highly divergent families and each family has its own multiple features that co-evolved into FAEs along with specificity. Previous research showed that not all homologous proteins have analogous functions (Benner et al, 2000) and proteins sharing promiscuous domains are known to have different functions (Marcotte et al, 1999). The presence of a common domain with Ser-His-Asp catalytic triad within different FAEs does not imply that they have the same function and can act on the same substrates.

Sequence-derived descriptor features can represent and distinguish proteins with different functional and interaction profiles irrespective of sequence similarity (Han et al 2004). Every enzyme sequence can be represented by its respective descriptor vectors from encoded representations of twenty amino acid residues (Cai et al, 2004). One of the latest applications of machine learning is the successful use of physicochemical properties and sequence derived descriptors for the classification of G-protein coupled receptors (Karchin et al, 2002), nuclear receptors (Bhasin and Raghava, 2004) and for the subcellular localization of bacterial proteins (Bhasin et al, 2005). The efficacy of protein descriptors in the prediction of protein functional families that includes G protein-coupled receptors, transporter TC8.A, chlorophyll, proteins involved in lipid synthesis, and rRNA-binding proteins has been also described recently (Ong et al, 2007). We believe that by selecting and combining descriptors that contain complementary type of information related to protein-ligand binding, one can enhance the performance of protein family classifiers. The initial step for the present classification system of the putative and known FAEs was the unsupervised clustering of sequences based on a large number of sequence derived descriptors. Following this, a machine learning algorithm was trained to predict the class of new FAEs.

Sequence-based descriptor sets

Our dataset contains 365 characterized, partially characterized and putative FAE sequences retrieved and filtered as described in Section 2. For each FAE protein sequence, descriptors were generated using PROFEAT - Protein Feature Server (Li et al, 2006) with the exception of physicochemical composition descriptors that were generated using the COPid server (Kumar et al, 2008). The algorithms for generating respective sequence derived descriptors applied in our study are briefly described below.

Amino acid composition: The amino acid composition descriptor set represents the occurrence frequency of all natural amino acids in a protein sequence. A total of 20 descriptor values were computed for the 20 types of amino acids i.e., this descriptor set corresponds to a 20 dimensional feature vector. Amino acid composition is defined as the fraction of each amino acid type in a protein sequence

$$f(a) = \frac{N_a}{N} \quad [1]$$

where $a = 1, 2, 3, 4, 5 \ldots 20$

N_a = number of amino acid of type a

N = length of the protein sequence

Dipeptide composition: Dipeptide composition represents the occurrence frequency of all consecutive amino acid pairs (20×20=400) in a protein sequence and corresponds to a 400 dimension feature vector. In contrast to amino acid composition, this descriptor set can encapsulate information about composition of amino acids as well as their local order. The combination of amino acid composition and dipeptide composition descriptor sets has been successfully used by researchers for classification of nuclear receptors (Bhasin and Raghava, 2004) and classification of G-coupled receptors (Gao and Wang, 2006). Dipeptide composition is defined as

$$f(a,b) = \frac{N_{ab}}{N-1} \quad [2]$$

where $a,b = 1, 2, 3, 4, 5 \ldots 20$

N_{ab} = number of dipeptides composed of amino acid type a and b

Autocorrelation descriptors: Autocorrelation descriptors are defined based on the distribution of amino acid properties along the sequence, also known as molecular connectivity indices and belong to a class of topological descriptors that describe the level of correlation between two objects in terms of their specific structural or physicochemical property (Broto et al, 1984). The different amino acid properties used as autocorrelation descriptors were various types of amino acids indices viz., hydrophobicity scales (Cid et al, 1992), average flexibility indices (Bhaskaran and Ponnuswamy, 1988), polarizability parameter (Charton M and Charton BI, 1982), free energy of solution in water (Charton M and Charton BI, 1982), residue accessible surface area in tripeptide (Chothia, 1976), residue volume (Bigelow, 1967), steric parameter (Charton, 1981) and relative mutability (Dayhoff et al, 1979). Three different autocorrelation descriptors viz., Moreau-Broto autocorrelation, Moran autocorrelation and Geary autocorrelation descriptors are computed, each having 240 descriptor components.

Moreau-Broto autocorrelation uses the property values as the basis of measurement, whereas Moran autocorrelation utilizes property deviations from average values, and Geary autocorrelation descriptors measure spatial autocorrelation (correlation of a variable with itself through space). In the past, autocorrelation descriptors have been successfully implemented for prediction of membrane protein types (Feng and Zhang, 2000), prediction of protein structural content (Lin and Pan, 2001) and for prediction of protein helix content (Horne, 1988).

All of the amino acid indices are centralized and standardized, i.e.

$$P'_a = \frac{P_a - \bar{P}}{\sigma} \quad [3]$$

where = average value of a particular property of the 20 amino acids

and are given by :

$$\bar{P} = \frac{\sum_{a=1}^{20} P_a}{20} \quad [4]$$

$$\sigma = \sqrt{\frac{1}{20}\sum_{a=1}^{20}(P_a - \bar{P})^2} \quad [5]$$

Moreau-Broto autocorrelation descriptors are defined as:

$$AC(d) = \sum_{i=1}^{N-d} P_i P_{i+d} \quad [6]$$

where d = 1, 2, 3, 4, 5 ...30 is the lag of the autocorrelation

P_i and P_{i+d} = amino acid property at position i and $i+d$ respectively

The normalized Moreau-Broto autocorrelation is defined as:

$$ATS(d) = \frac{AC(d)}{N-d} \quad [7]$$

The Moran autocorrelation algorithm uses the property deviations from the average values and is defined as:

$$I(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}(P_i - \bar{P})(P_{i+d} - \bar{P})}{\frac{1}{N}\sum_{i=1}^{N}(P_i - \bar{P})^2} \quad [8]$$

where d, P_i and P_{i+d} are defined same as above.

$\bar{P}$ = average of the considered property P along the sequence, i.e.

The Geary autocorrelation algorithm uses square difference of property values and is defined as:

$$C(d) = \frac{\frac{1}{2(N-D)}\sum_{i=1}^{N-d}(P_i - P_{i+d})^2}{\frac{1}{N-1}\sum_{i=1}^{N}(P_i - \bar{P})^2} \quad [9]$$

where d, $\bar{P}$, P_i and P_{i+d} are defined in the same way as above.

Descriptors of composition (C), transition (T) and distribution (D): CTD descriptors represent the amino acid distribution patterns of a particular physicochemical property along the protein sequence. These descriptors, developed by Dubchak et al (1995), have been successfully used for functional classification of proteins from their primary sequence (Cai et al, 2003b), predicting functional family of enzymes (Han et al, 2004; Lin et al, 2006) and for prediction of protein folding class (Dubchak et al, 1999). CTD descriptors comprise attributes of seven structural or physicochemical properties and each attribute is further divided into three groups as shown in Table 3.

Table 3 Amino acid attributes and further division of each attribute into three groups of amino acid residues.

Attribute	Divisions		
Secondary structure	Helix	Strand	Coil
	EALMQKRH	VIYCWFT	GNPSD
Solvent accessibility	Buried	Exposed	Intermediate
	ALFCGIVW	PKQEND	MPSTHY
Charge	Positive	Neutral	Negative
	KR	ANCQGHILMFPSTWYV	DE
Hydrophobicity	Polar	Neutral	Hydrophobicity
	R,K,E,D,Q,N	G, A, S,T,P,H,Y	C,L,V,I,M,F,W
Polarity	Polarity value 4.9–6.2	Polarity value 8.0–9.2	Polarity value 10.4–13.0
	L,I,F,W,C,M,V,Y	P,A,T,G,S	H,Q,R,K,N,E,D
Polarizability	Polarizability value 0–1.08	Polarizability value 0.128–0.186	Polarizability value 0.219–0.409
	G,A,S,D,T	C,P,N,V,E,Q,I,L	K,M,H,F,R,Y,W
Normalized van der Waals volume	Volume range 0–2.78	Volume range 2.95–4.0	Volume range 4.03–8.08
	G,A,S,T,P,D	N,V,E,Q,I,L	M,H,K,F,R,Y,W

Three values of Composition (C) descriptors are then computed for a given attribute to describe the global percent composition of the three groups along the protein sequence. For example, hydrophobicity attribute consists of three values viz., the global percent compositions of polar, neutral and hydrophobic residues. Transition (T) descriptors also consists three values for a given attribute and are computed as the percent frequencies with which the attribute changes its index along the entire length of the protein. Distribution (D) descriptors consist of five values for each of the three groups of a given attribute and are computed as the distribution pattern of the attribute along the protein sequence. So, there are 3 descriptors and 3(C) + 3(T) + 5×3(D) = 21 descriptor values for each attribute. In total the seven attributes produce a total of 7 × 3 = 21 descriptors and the final protein feature vector of the CTD descriptor set contains 7 × 21 = 147 descriptor values.

Sequence-order descriptors: Sequence-order descriptors proposed by Chou (Chou, 2000), count the physicochemical distance between amino acids. The physicochemical properties computed include hydrophobicity, polarity and side chain volume. For each amino acid type, a sequence-order descriptor is derived from both the Schneider-Wrede physicochemical distance matrix (Schneider and Wrede, 1994) and the normalized Grantham chemical distance matrix (Grantham, 1974). For a protein sequence of N amino acid residues, the sequence order effect can be reflected through a set of sequence order coupling

numbers. The dth rank sequence order-coupling number (d = 1, 2, 3, 4, 5 ...30) is defined as:

$$\tau_d = \sum_{i=1}^{N-d} \left(d_{i,i+d}\right)^2 \quad [10]$$

where $d_{i, i+d}$ = physicochemical distance between the two amino acids at position i and $i+d$

For each amino acid type, the first 20 quasi-sequence-order descriptors are defined as:

$$Xr = \frac{f_r}{\sum_{r=1}^{20} f_r + w \sum_{d=1}^{30} \tau_d} \quad [11]$$

where r = 1, 2, 3, 4, 5 ...20

The other 30 quasi-sequence-order descriptors are defined as:

$$Xd = \frac{w\tau_{d-20}}{\sum_{r=1}^{20} f_r + w \sum_{d=1}^{30} \tau_d} \quad [12]$$

where d = 21, 22, 23, 24, 25 ...50

There are 30 descriptor components for sequence-order-coupling numbers derived from Schneider-Wrede physicochemical distance matrix and another set of 30 descriptor components for sequence-order-coupling numbers derived from normalized Grantham chemical distance matrix. Quasi-sequence-order descriptor set contains 50 descriptor components for each type of distance matrix as described above. These descriptors have been used for predicting protein subcellular locations (Chou and Cai, 2004).

Descriptors of pseudo amino acid composition: This descriptor set is a combination of amino acid composition with 20 dimensional vectors and another 30 dimensional vector reflecting sequence order correlated factors (Chou, 2001). This descriptor set has been successfully used for prediction of enzyme subfamily classes (Chou, 2005).

Physicochemical composition: This set contains 11 descriptor components viz., percentage compositions of charged, aliphatic, aromatic, polar, neutral, hydrophobic, positive charged, negative charged, tiny, small and large residues.

Clustering of FAEs and putative FAEs

Machine learning approaches help us gain knowledge from complex patterns in data. Clustering is an unsupervised machine learning technique that reveals how instances are naturally grouped in the descriptor space. In clustering, the classes are unknown and are identified by the cluster analysis of the data. In simple terms, the overall idea of clustering is to group similar elements together.

Clustering studies on FAEs described in this study were performed using J-Express 2009 Package - Version 1.3 (J-Express, Molmine AS, Norway, http://www.molmine.com/). An intermediate between clustering and multidimensional scaling is provided by J-Express (Dysvik and Jonassen, 2001) through the implementation of a Self-Organizing Map (SOM) algorithm. A feature in the latest version 1.3 of J-Express 2009 package that allows the user to intitate k-means clustering from the clusters resulted by SOM was applied, which overcomes the problem of having to specify the number of clusters in traditional k-means clustering.

The principal feature of SOM is the 2-dimensional rendering of a multidimensional space, which brings similar instances in close vicinity within the same neuron on the map (Kaiser et al, 2007). The SOM algorithm starts by placing each neuron in the input space by giving it a reference vector equal to an arbitrary input value (Kohonen, 2001). Each SOM iteration step consists of randomly selecting a point from the input data set and moving the nearest node (winner node) and its neighborhood towards it (Garrigues et al, 2005). A neighbourhood function is used to determine the learning rate and the rate of change around the winner node, i.e. how much the node and its neighborhood will move in the direction of the input vector. SOM training results in transforming the lattice into an 'elastic surface' that is stretched over the input. At the end, each cluster is defined with reference to a node, specifically comprised by those data points for which it represents the winner node.

For the training of SOM, we used a starting input grid of 100 neurons and 4000 iterations, while the Gauss neighborhood function (Lee and Verleysen, 2002) and the Euclidean distance measure (Oili and Markku 2008) were applied for updating the grid. Subsequently, the output of SOM was fed as input vector to the k-means clustering algorithm, in order to define the boarders between the nodes and to put in the same cluster nodes that were close to each other. For consistency, we used again 4000 iterations and Euclidean distance measure. The K-means algorithm is an iterative two-step algorithm. In the first step, each data point is assigned to the nearest mean. In the second step, the means are adjusted to match the sample means of the data points that they are responsible for (MacKay and David, 2003). K-means clustering of SOM readily identifies similar patterns in data. The approach of K-means clustering of SOM has been successfully employed by researchers for the identification of distinct gene expression patterns (Wang et al, 2002).

Clustering performance of different sequence-based descriptors

With the intention to select the best descriptor set that clusters FAEs with low variance within clusters and high variance between clusters, we evaluated the effectiveness of different descriptor sets listed in Table 4 as well as combinations of the ones showing the highest performance. The evaluation of the formed clusters was based on inspection of the within and between clusters variance. Clusters with low within variance and high variance between them, is what characterizes a good clustering output. Within and between clusters variance scores implemented in J-Express are according to Dudoit et al (2000). Table 4 summarizes the performance of the individual descriptor sets and their combinations.

Table 4 A summary of analysis on efficacy of different sequence derived descriptors. Based on the variance scores within and between clusters, descriptor set combination DS14 was chosen as the best set.

Set code	Descriptor sets	Descriptor Components	Number of clusters	Variance score	
				Within Cluster	Between Clusters
DS1	Amino acid composition	20	12	0.001	6.62
DS2	Dipeptide composition	400	12	0.001	0.07
DS3	Normalized Moreau-Broto autocorrelation descriptors	240	17	8.92	0.005
DS4	Moran autocorrelation descriptors	240	23	1.86	0.001
DS5	Geary autocorrelation descriptors	240	13	2.91	0.001
DS6	Composition, transition, distribution	147	13	2755	1024
DS7	Sequence order coupling numbers (Schneider-Wrede physicochemical distance matrix)	30	13	392027	14.4
DS8	Sequence order coupling numbers (Grantham chemical distance matrix)	30	13	16499	8.52
DS9	Quasi sequence order descriptors (Schneider-Wrede physicochemical distance matrix)	50	12	0.001	0.001
DS10	Quasi sequence order descriptors (Grantham chemical distance matrix)	50	13	0.001	0.001
DS11	Pseudo amino acid composition	50	10	0.001	0.001
DS12	Physico-chemical composition	11	12	16.14	147.57
DS13	Amino acid composition and dipeptide composition	420	12	0.001	1.42
DS14	Amino acid composition and physico-chemical composition	31	13	14.15	157.44
DS15	Dipeptide composition and physico-chemical composition	413	11	15.44	20.54
DS16	Amino acid composition, dipeptide composition and physico-chemical composition	433	12	13.13	20.53

Descriptors of amino acid composition (DS1), dipeptide composition (DS2) and physicochemical composition (DS12) showed satisfactory variance scores within and between clusters. On the other hand, the rest of the descriptor sets (DS3, DS4, DS5, DS6, DS7, DS8, DS9, DS10 and DS11) showed poor performance with low quality variance scores. Furthermore, combination of well-performing descriptors sets with complementary information improve further the clustering of data (as evident from Set DS14 that results in FAE clusters with better within and between variance compared to the individual descriptor sets). The combination of the amino acid composition and physico-chemical composition descriptor sets (DS14) outperformed all other sets (DS13, DS15 and DS16) in the clustering of putative and known FAE-sequences. The members of each cluster along with the source organism are listed in Table S9 of Supplementary File 2 and are summarized in Table 5. In order to evaluate the clustering in terms of biochemical relevance we investigated the distribution of the previously characterized FAEs. Three FAEs previously characterized as type A, from *Aspergillus awamori*, *Aspergillus niger* and *Aspergillus tubingensis* were clustered in cluster 13, together with one FAE known as type B from *Aspergillus oryzae*. Another three of the FAEs type B from *Aspergillus nidulans*, *Penicillium chrysogenum* and *Aspergillus niger* were clustered in cluster 4, together with three type C FAEs from *Aspergillus oryzae, Talaromyces stipitatus* and *Fusarium oxysporum*; this cluster also accommodates a type D FAE from *Neurospora crassa*. Two type B FAEs from *Penicillium funiculosum* and *Neurospora crassa* were clustered each in cluster 5 and cluster 6, respectively. A type D FAE from *Piromyces equi* is the only member of cluster 2, and is a modular cinnamoyl ester hydrolase of a multiprotein cellulose-binding cellulase-hemicellulase complex (Fillingham et al, 1999). None of the characterized FAEs were present in the remaining clusters.

Table 5 The thirteen FAE clusters obtained from descriptor set DS14. The size of each cluster and the distribution of FAEs characterized as type-A, B, C and D among the clusters is shown.

Cluster	No of sequences	Distribution of type A, B, C and D FAEs
C1	35	-
C2	1	Type D FAE
C3	30	-
C4	51	3 type B FAEs, 3 type C FAEs, 1 type D FAE
C5	17	1 type B FAE
C6	29	1 type B FAE
C7	49	-
C8	34	-
C9	18	-
C10	18	-
C11	30	-
C12	29	-
C13	24	3 type A FAEs, 1 type B FAE

Training of a Support Vector Machine model for the classification of FAEs

The goal of clustering is to group data based on common traits, whereas classification deals with the assignment of an unknown instance to a specific class among a predefined number of classes (Gasteiger and Engel, 2003). Support vector machines (SVM) are supervised learning methods that learn by example to assign labels to objects (Noble, 2006) and perform the classification by constructing an *N*-dimensional hyperplane that optimally separates the data with different labels.

The training and optimization of a SVM classifier was performed in WEKA (Waikato Environment for Knowledge Analysis), a java software package from University of Waikato, New Zealand (http://www.cs.waikato.ac.nz/ml/weka). Sequential Minimal Optimization (SMO) algorithm was used for training a SVM classifier (Platt, 1998; Keerthi et al, 2001). Four different kernel functions viz., linear, polynomial, RBF and sigmoid kernels were evaluated. SVM classifiers have been extensively described in the literature (Chang and Lin, 2001, Uestuen et al, 2006; Hsu et al, 2009) for protein functional classification (Cai et al, 2003a), enzyme family classification (Cai et al, 2004) and for protein secondary structure prediction (Kim and Park, 2003). The model was trained using 10-fold cross-validation. Cross-validation helps for assessing how the results of a statistical analysis will generalize to an independent data set (Kohavi, 1995). In 10-fold cross-validation, the data set is divided into ten subsets, and the holdout method is repeated ten times. Each time, one of the ten subsets is used as the test set and the other (10–1) subsets are put together to form a training set. Then the average error across all ten trials is computed.

The performance of different kernels and the respective parameters on the classification process of the FAE clusters was evaluated and the best SVM model was further validated against a blind test set. The blind test set consisted of 37 sequences that were selected by randomly removing from the original data set the 10% of the sequences of each cluster prior to the training of the model. The summary of the best SVM model selected for classification of FAE clusters obtained using the descriptor set DS14 is given in Table 6.

The SVM model successfully recognized 98.2% of the sequences that belong to respective clusters and all the sequences of the blind test set. The consistency of the performance in both training and blind test sets, demonstrates the validity of the present method. The high quality of the SVM model developed here guarantees the correct classification of any new FAE sequence that will arise from genome-sequencing projects or improved annotation algorithms

Table 6
Performance of best SVM model for classification of FAEs and putative FAEs

SVM Type: Sequential Minimal Optimization (SMO); Kernel Type: Radial Basis Function (RBF) kernel											
Training set performance						Validation with blind test set					
Class	TPR [a]	FPR [b]	ROC Area [c]	% of Correctly Classified Instances	% of Incorrectly Classified Instances	Class	TPR [a]	FPR [b]	ROC Area [c]	% of Correctly Classified Instances	% of Incorrectly Classified Instances
C1	1	0	1			C1	1	0	1		
C2	1	0	1			C2	-	-	-		
C3	1	0.003	0.998			C3	1	0	1		
C4	1	0	1			C4	1	0	1		
C5	0.933	0	0.997			C5	1	0	1		
C6	1	0	1			C6	1	0	1		
C7	1	0.014	0.993			C7	1	0	1		
C8	0.968	0.003	0.997	98.2%	1.8%	C8	1	0	1	100%	0%
C9	1	0	1			C9	1	0	1		
C10	1	0	1			C10	1	0	1		
C11	0.963	0	0.998			C11	1	0	1		
C12	0.962	0	0.999			C12	1	0	1		
C13	0.909	0	0.997			C13	1	0	1		
Weighted Average	0.982	0.002	0.998			Weighted Average	1	0	1		

[a] True Positive Rate (TPR): The True Positive rate or Sensitivity is the proportion of examples which were classified as class *'A'*, among all examples which truly have class *'A'*.
[b] False Positive Rate (FPR): The False Positive rate or 1–[Specificity] is the proportion of examples which were classified as class *'A'*, but belong to a different class, among all examples which are not of class *'A'*.
[c] ROC Area: A Receive Operating Characteristic (ROC) is a graphical plot of the sensitivity *vs.* (1 – specificity)

Biological basis of the FAE classification system

The goal of classification is to group together functionally related FAEs that have common properties. To assign protein superfamilies to all 365 FAE-related sequences, we have used the UFO server (Meinicke, 2009) that provides a fast detection of protein domains according to the Pfam A release 23 which comprises 10340 domain families (Finn et al, 2008). UFO also contains the precomputed profiles of 821 genomes that comprise 54 archaeal, 721 bacterial and 46 eukaryotic proteomes respectively from the HAMAP database (Lima et al, 2009) which are used for profile comparison. The complete list of superfamilies and respective probability scores for FAE-related sequences were given in the Supplementary File 3 (Table S10) and are summarized in Figure 4. The probability score for a superfamily is in the range between 0.5 and 1.0 with high values above 0.9 usually indicating good match. Out of 365 sequences the majority presented a probability score higher than 0.98, however, twenty sequences do not have an assignment of a superfamily. From the rest, 237 sequences belong to tannase and feruloyl esterase protein superfamily, 41 sequences belong to putative esterase protein superfamily, 31 sequences belong to esterase PHB depolymerase protein superfamily and 10 sequences belong to lipase (class 3) protein superfamily.

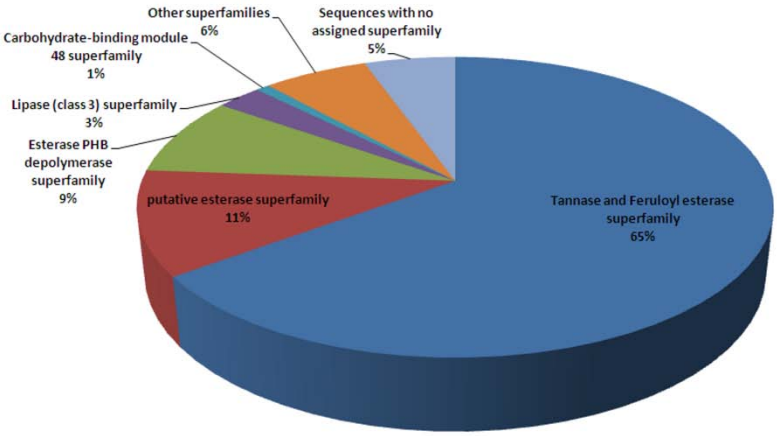


Figure 4 Pie plot showing the distribution of FAE-related sequences among the protein superfamilies. Details of each sequence along with probability scores were given in Table S10 (Supplementary File 3).

To find out whether the sequences that are predicted to fall in the same cluster share some common domains of highly conserved regions, we used Block Maker tool (Henikoff et al, 1995) that finds blocks in a group of related protein sequences. Interestingly, we found that each cluster has its respective pattern of blocks or domains that the sequences share. Blocks graphical map of respective FAE clusters and the positions of domains in each sequence of respective conserved block are given in Supplementary File 3 (Table S11). The number and length of blocks that mark each cluster varied significantly between the clusters. For example, cluster 7 contains seven blocks (Figure 5a to 5g) with the longer being 35 amino acids (block 4) and the shorter to have 12 amino acids length (block 6). In contrast, cluster 5 is characterized by only two blocks of 17 and 11 amino acid lengths respectively (Figure 5h and 5i). The sequence logo diagrams were created using LogoBar (Pérez-Bercoff et al, 2006).

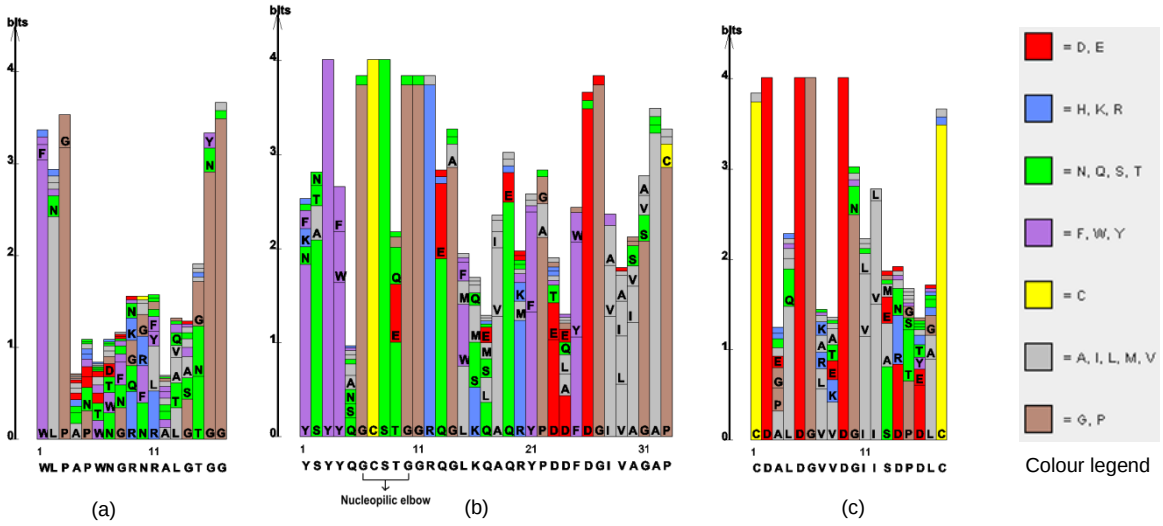
Interestingly, FAE-related sequences that belong to different protein superfamilies share conserved blocks and were grouped together as different clusters (Supplementary File 3 Table S10 & Table S11). For example, in cluster 10, 61% of sequences belong to tannase-feruloyl esterase superfamily and 28% sequences belong to putative esterase superfamily; where as 56% and 44% sequences of this cluster were fungal and bacterial origin respectively. The conservation of two superfamily blocks in both bacteria and fungi supports the complex evolutionary relationship between the FAE-related sequences of different superfamilies among the kingdoms as discussed in Section 2.

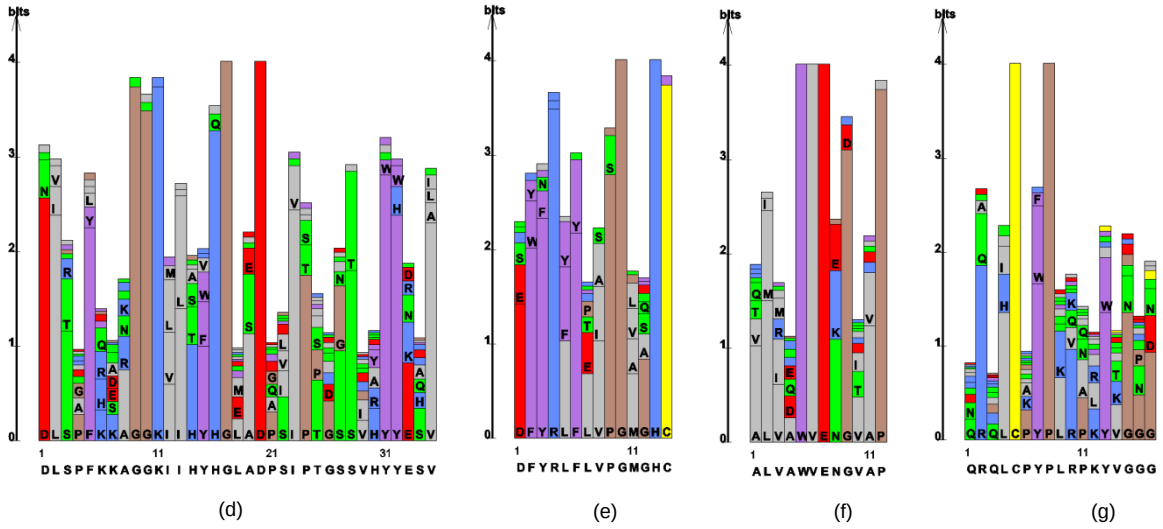
Analysis of the available FAE sequences and available crystal structures by several researches shows that the active site of FAEs contains the catalytic triad (Ser, His, Asp), and the serine residue is located at the centre of universally conserved pentapeptide with the consensus 'nucleophilic elbow' i.e., GXSXG (X=any amino acid residue) (Schubot et al, 2001; Prates et al, 2001; Tarbouriech et al, 2005; Hermoso et al, 2004; McAuley et al, 2004; Faulds et al, 2005; Benoit et al, 2006b). So, the presence of the catalytic triad with the serine containing nucleophilic elbow in a particular candidate sequence denotes a high probability for it being a putative FAE. We analyzed all the candidate sequences for the presence of 'nucleophilic elbow' using the BioEdit program (Hall, 1999). Interestingly some of the candidate sequences showed multiple 'nucleophilic elbows' which led us to the hypothesis that these proteins may have more than one functional site or binding pockets (Supplementary File 3 – Table S12).

Predicting the functional residues of the catalytic triad of every candidate sequence is very important, as this will help in filtering out sequences but also for further sub-classification of the FAEs. The distance between the catalytic residues boosts the accuracy of the sub-classification system of FAEs and highlights spatial patterns of conservation reflecting the structural information. To predict functionally important residues, a recently developed tool, INTREPID (Sankararaman et al, 2009) was used, which computes an information-theoretic score for each position in the sequence. INTREPID uses Jensen-Shannon divergence to measure the information for each position in the sequence at each sub-tree node encountered on a traversal of the phylogeny, tracing a path from the root to the leaf corresponding to the sequence of interest (Sankararaman and Sjölander, 2008). The advantage of INTREPED server is that it makes full use of the information in a protein family containing many distantly related sequences through use of tree traversal with the ability to detect subtle evolutionary patterns that other prediction methods might miss. The catalytic triad residues of respective candidate FAEs are given in Supplementary File 3 (Table S13). The INTREPID prediction results were cross-checked with published experimental data for available FAEs and found to be very accurate. FAE from *Penicillium funiculosum* was experimentally studied and 118 Ser, 202 Asp, 258 His were predicted as the catalytic triad (Kroon et al, 2000); INTREPED accurately predicted 118 Ser, 202 Asp, and 258 His as catalytic triad (see Table S13 of Supplementary File 3).

INTREPED accurately predicts the distance between catalytic residues of *Clostridium thermocellum* FAEs, as evident from the crystal structures of FAE_XynZ and FAE_XynY (Blum et al, 2000; Schubot et al, 2001; Prates et al 2001). In case of FAE (AnFaeA) from *Aspergillus niger*, INTREPED was able to correctly predict the distance between Ser and Asp; the prediction score for His132 as a functional residue is high when compared with score of His247 in contrary to the analysis done by McAuley et al, 2004. To maintain consistency, we strictly considered the output from INTREPED for prediction of catalytic triad residues in sequences used for this sub-classification system.

Out of 365 putative FAE sequences, 41 sequences do not contain the nucleophilic elbow or catalytic triad residues and those sequences were filtered out. The remaining 324 sequences of the respective clusters were classified into sub-groups based on the distance between the catalytic triad residues. By removal of sequences of cluster 9 that were not related to FAEs, the 13 clusters are reduced to 12, and are referred to as FEF1-12 (Feruloyl Esterase Family) from this point on. Based on the constellation and distance between the catalytic residues (S, D, H), the 12 FEFs were sub-grouped. For example, the family 11 (FEF11) was sub-grouped into 11A and 11B. In sub-group 11A, the average distances between catalytic triad residues were [Serine -228- Aspartic acid -38- Histidine], whereas the average between catalytic triad residues in sub-group 11B were [Serine -18- Aspartic acid -207- Histidine]. We analyzed the distribution of characterized fungal FAEs that have been classified into A, B, C and D types by Crepin et al (2004) among sub-families of the proposed FEFs. A, B, C and D FAE types maintained certain pattern in the distances between catalytic triad residues and fall in different sub-families (Supplementary File 4 – Table S14). All type A FAEs cluster together in sub-family FEF12A of our classification system. The FAE sequences of *Aspergillus nidulans, Penicillium chrysogenum,* and *Aspergillus niger* which have been characterized as type B, fall in our classification system at the sub-family FEF4C, however other type B sequences from *Penicillium funiculosum, Neurospora crassa* and *Aspergillus oryzae* fall into sub-families FEF5B, FEF6A and FEF12B, respectively. On the other hand all type C FAEs cluster together at the sub-family FEF4B. For several of our proposed families there are no members that have been experimentally characterized. FEF3 and FEF7





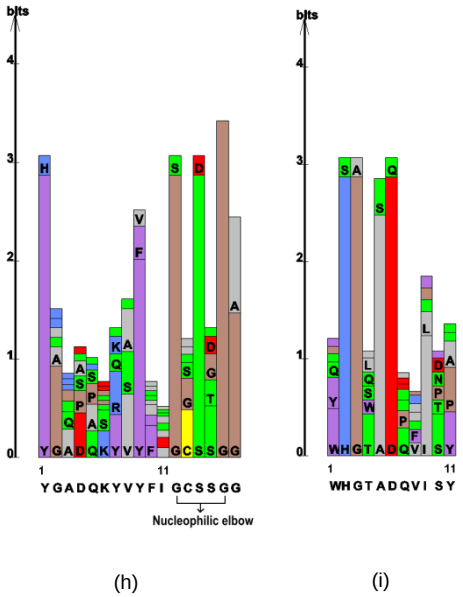


Figure 5 Protein sequence logo diagrams of conserved blocks created using LogoBar (Pérez-Bercoff et al, 2006). (a), (b), (c), (d), (e), (f) and (g) are the seven conserved blocks in the sequence members of cluster 7. (h) and (i) are the two conserved blocks in the sequence members of cluster 5. Within each bar, the letter for the corresponding amino acid is shown. The most frequent amino acids in each block were given at the bottom of respective protein sequence logo diagram. The conserved blocks containing the nucleophilic elbow (GXSXG) is indicated in the figure. Colour legend shows the colour-grouping of the amino acids. Red: polar negative amino acids (D, E); Blue: polar positive amino acids (H, K, R); Green: polar neutral amino acids (N, Q, S, T); Purple: non-polar aromatic amino acids (F, W, Y); Yellow: Cysteine (C); Grey: non-polar aliphatic amino acids (A, I, L, M. V); Brown: Glycine and Proline (G, P).

contain FAE sequences dominated by gram negative bacteria and fungi, respectively. All the other families accommodate a mixture of sequences from fungi, bacteria and plantae, which signifies that FAE-related sequences might have co-evolved together from a common ancestor that arose into different families during evolution of the respective kingdoms. The Table S14 (Supplementary File 4) shows the new proposed classification system of FAEs covering the fungal, bacterial and plantae kingdoms and the Table 7 shows the number of sequences falling into each family (FEF1-FEF12).

Table 7 The proposed classification system for feruloyl esterases from fungi, bacteria and plantae. The number of sequences in each family and the distribution of FAEs characterized as type-A, B, C and D among the FAE sub-families is shown.

Feruloyl esterase Family	No of sequences	Sub-group	No of sequences	Distribution of type A, B, C and D FAEs
FEF 1	33	1A	26	-
		1B	7	-
FEF 2	1	-	1	One Type D FAE
FEF 3	28	3A	24	-
		3B	3	-
		3C	1	-
FEF 4	50	4A	35	Three Type B FAEs
		4B	9	Three Type C FAEs
		4C	4	-
		4D	2	One Type D FAE
FEF 5	14	5A	7	-
		5B	5	One Type B FAE
		5C	2	-
FEF 6	27	6A	11	One Type B FAE
		6B	16	-
FEF 7	47	7A	36	-
		7B	9	-
		7C	2	-
FEF 8	33	8A	26	-
		8B	2	-
		8C	2	-
		8D	3	-
FEF 9	16	9A	8	-
		9B	6	-
		9C	2	-
FEF 10	23	10A	17	-
		10B	4	-
		10C	2	-
FEF 11	28	11A	25	-
		11B	3	-
FEF 12	24	12A	7	Three Type A FAEs
		12B	17	One Type B FAE

Structural analysis and substrate specificity of the new FAE families

The enzyme-substrate complex is formed when a substrate (the ligand) binds to the active-site pocket of the enzyme (the receptor). Typically ligands possess 3 to 15 rotatable bonds, whereas receptors possess many more (Teodoro and Phillips, 2001). These rotatable bonds give rise to 'degree of freedom' for the molecules. The rotatable bonds of active site residues and rotatable bonds of ligand in the enzyme-product complex of FAE – ferulic acid (PDB_ID: 1UWC) are shown in Figure 6(a). The major reason responsible for inaccuracies in the docking methods used at most of academic and industrial research is to assume a rigid protein. Several docking programs and methods assume that proteins (receptors) are rigid macromolecules and the respective substrates (ligands) during the binding process changes their three dimensional structure for a best spatial and energetic fit in the receptor's site (Morris et al, 1998; Ewing and Kuntz, 1998).

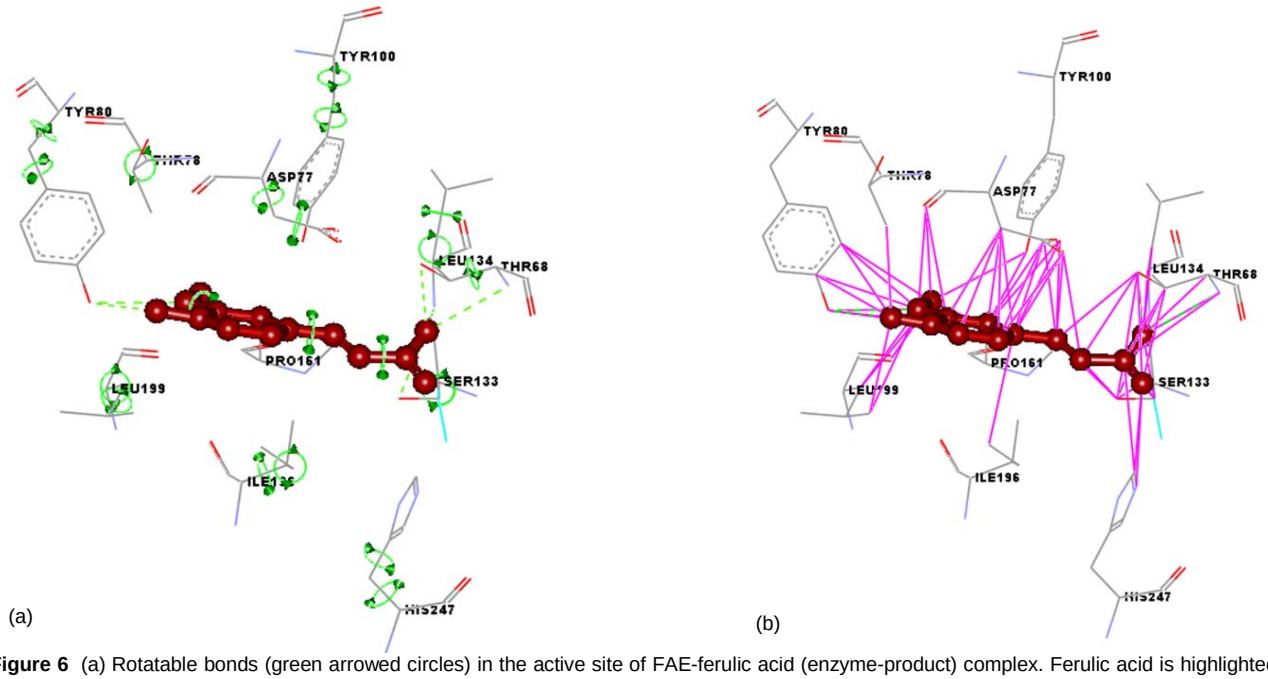


Figure 6 (a) Rotatable bonds (green arrowed circles) in the active site of FAE-ferulic acid (enzyme-product) complex. Ferulic acid is highlighted in red with ball and stick display style. (b) Ligand binding pattern calculated using Accelrys Discovery Studio Client 2.5 (2009), that displays any hydrogen bonds (green dashed lines), close interactions like pi-pi, pi-cation and pi-sigma interactions found between the protein and the ligand molecule (magenta lines).

The alternative binding modes can be explored if several X-ray structures of protein-ligand complexes are available. FAE to a resolution of 1.50Å (PDB_ID: 1UZA) and FAE (from *Aspergillus niger*) in complex with its product ferulic acid to a resolution of 1.08Å (PDB_ID: 1UWC) has been determined by McAuley et al in 2004. Two years later, Benoit et al (2006b) have determined the structure of FAE in complex with CAPS (cyclohexyl amino propane sulfonic acid) to a resolution of 1.55Å (PDB_ID: 2HL6). Analysis of ligand binding site of FAE-FA complex (1USW-A chain) showed that three water molecules are present within the distance of 4Å. The changes in ligand binding site residues of the FAE are shown in Figure 7, indicating the flexibility of the receptor.

From the surface view diagrams of Figure 7, it is evident that only the feruloyl moiety is present in the vicinity of active site residues in the binding pocket and the position of CAPS will be taken by the sugar residues of natural substrates (Benoit et al, 2006b). Researchers have shown that tight binding of carbohydrate moiety is not required for catalysis (Schubot et al, 2001) and the specificity of FAE wholly depends upon the type of phenolic acid to which the carbohydrate moiety is attached by an ester bond. Faulds et al (2005) crystallized the inactive S133A mutant of FAE in complex with a feruloylated trisachharide substrate, but they found that only the ferulic acid moiety of the substrate is visible in the electron density map, which is in agreement with that the carbohydrate moiety does not play a major role in the binding. In simple terms, we can say that FAE recognizes the cinnamoyl structure of the substrate, while the esterified carbohydrate residue is less important for binding of substrate in the active site. As we mentioned earlier, the essentiality of a classification system is to group functionally related FAEs. Only seven FAEs have been experimentally tested on a wide range of substrates (Topakas et al, 2004; Vafiadi et al, 2006) and amino acid sequences of three of them (AnFeaA, AnFaeB and TsFaeC) are known. Information on the activity of 25 substrates for these three FAEs and their respective specificities is available from literature (Kroon et al, 1997; Topakas et al, 2004; Vafiadi et al, 2006) and is summarized in Table 8.

We cross-checked the distribution of the three FAEs (AnFaeA bearing gi|17366177|sp|O42807.1|; AnFaeB bearing gi|17932783|emb|CAC83933.1|; TsFaeC bearing gi|33945411|emb|CAD44531.1|) among our proposed classification system. We found that AnFaeA falls under the sub-family 12A of the FEF12 family, AnFaeB falls under the sub-family 4A of the FEF4 family and TsFaeC falls under the sub-family 4B of the FEF4 family (see Supplementary File 4 – Table S14). It should also be noted that these three FAEs have been classified previously as Type A, B and C, respectively, by Crepin et al (2004). Furthermore, all the three type A FAEs were clustered in sub-group 12A and all three type C FAEs were clustered in sub-group 4B of our FAE classification system. Thus, our framework is very successful in sub-grouping functionally related FAEs together based on sequence derived descriptors followed by focus on distance between catalytic triad residues. The pharmacophore analysis that follows, provide information on the pharmacophoric features of the substrates that are necessary to ensure the optimal supramolecular interactions with FAEs of specific sub-family. Due to limited substrate information for the other sub-families, we develop common feature-based pharmacophore models for three FAE sub-families 12A, 4A and 4B.

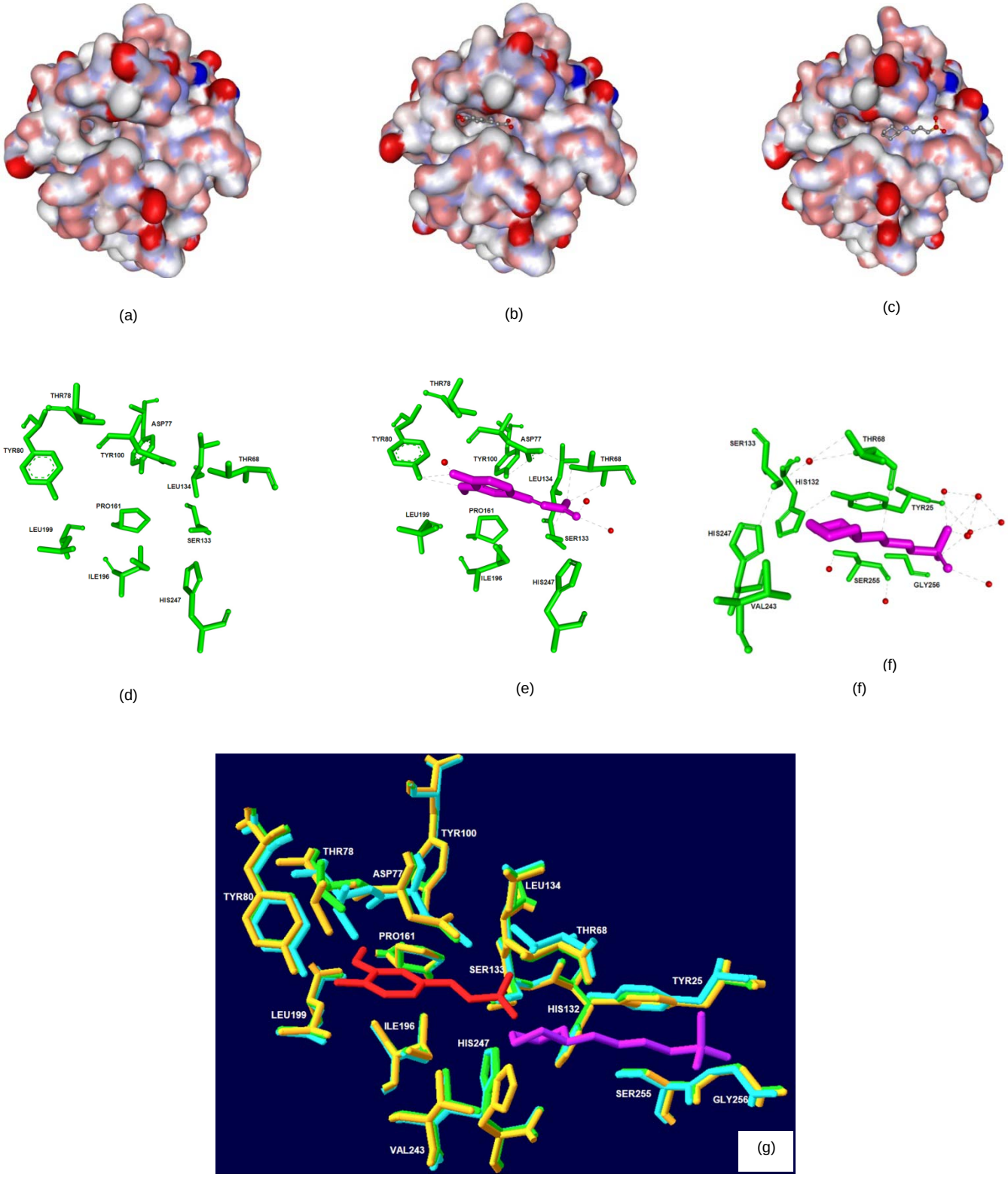


Figure 7 Flexibility of residues in the ligand binding site of *Aspergillus niger* FAE. (a) Surface view of FAE (McAuley et al, 2004). (b) Surface view of FAE-FA complex (McAuley et al, 2004). Ferulic acid is represented in ball and stick model. (c) Surface view of FAE-CAPS complex (Benoit et al, 2006b). CAPS is represented in ball and stick model. (d) Enlarged view of ligand binding site residues. (e) Ligand binding site residues of FAE-FA complex. (f) Ligand binding site residues of FAE-CAPS complex. (g) Aligned ligand binding site residues of FAE, FAE-FA complex, FAE-CAPS complex. Binding of a ligand causes major changes in the side chains of amino acid residues. The changes in overall protein surface (colored according to atom charge) can be observed in FAEs shown in figures a, b and c. Ligand binding site atoms of respective FAEs are shown in figures d, e and f. Hydrogen bonds are shown in dashed lines. Water molecules (in red) present within 4Å distance from respective ligands are also shown. Receptor structures were prepared using *Protein preparation wizard* in Maestro (Schrödinger Suite 2009).

Table 8 Substrate specificities (information on 25 substrates is available) of three partially characterized FAEs for which amino-acid sequences are known. Respective enzymes can act on substrates highlighted in bold. None of the enzymes have activity on the substrates highlighted in italics.

AnFaeA [gi|17366177|sp|O42807.1|]	AnFaeB [gi|17932783|emb|CAC83933.1|]	TsFaeC [gi|33945411|emb|CAD44531.1|]
Methyl 3-methoxy cinnamate	**Methyl cinnamate**	**Methyl cinnamate**
Methyl 3,4-dimethoxy cinnamate	**Methyl 3-hydroxy cinnamate**	**Methyl 2-hydroxy cinnamate**
Methyl 3,5-dimethoxy cinnamate	**Methyl 4-hydroxy cinnamate (or) Methyl p-coumarate**	**Methyl 3-hydroxy cinnamate**
Methyl 3,4,5-trimethoxy cinnamate	**Methyl 3,4-dihydroxy cinnamate (or) Methyl caffeate**	**Methyl 4-hydroxy cinnamate (or) Methyl p-coumarate**
Methyl 4-hydroxy-3-methoxy cinnamate (or) Methyl ferulate	**Methyl 2-methoxy cinnamate**	**Methyl 3,4-dihydroxy cinnamate (or) Methyl caffeate**
Methyl 4-hydroxy-3,5-dimethoxy cinnamate (or) Methyl sinapate	**Methyl 4-methoxy cinnamate**	**Methyl 3-methoxy cinnamate**
Methyl 4-hydroxy-3-methoxy phenyl propionate	**Methyl 4-hydroxy-3-methoxy cinnamate (or) Methyl ferulate**	**Methyl 3,4-dimethoxy cinnamate**
Methyl cinnamate	**Methyl 3-hydroxy-4-methoxy cinnamate**	**Methyl 3,5-dimethoxy cinnamate**
Methyl 2-hydroxy cinnamate	**Methyl 4-hydroxy-3-methoxy phenyl propionate**	**Methyl 4-hydroxy-3-methoxy cinnamate (or) Methyl ferulate**
Methyl 3-hydroxy cinnamate	Methyl 2-hydroxy cinnamate	**Methyl 3-hydroxy-4-methoxy cinnamate**
Methyl 4-hydroxy cinnamate (or) Methyl p-coumarate	Methyl 3-methoxy cinnamate	**Methyl 4-hydroxy-3,5-dimethoxy cinnamate (or) Methyl sinapate**
Methyl 3,4-dihydroxy cinnamate (or) Methyl caffeate	Methyl 3,4-dimethoxy cinnamate	**Methyl 4-hydroxy-3-methoxy phenyl propionate**
Methyl 2-methoxy cinnamate	Methyl 3,5-dimethoxy cinnamate	Methyl 2-methoxy cinnamate
Methyl 4-methoxy cinnamate	Methyl 3,4,5-trimethoxy cinnamate	Methyl 4-methoxy cinnamate
Methyl 3-hydroxy-4-methoxy cinnamate	Methyl 4-hydroxy-3,5-dimethoxy cinnamate (or) Methyl sinapate	Methyl 3,4,5-trimethoxy cinnamate
Methyl 3,4-dichloro phenyl propionate	*Methyl 3,4-dichloro phenyl propionate*	*Methyl 3,4-dichloro phenyl propionate*
Methyl 4-hydroxy phenyl acetate	*Methyl 4-hydroxy phenyl acetate*	*Methyl 4-hydroxy phenyl acetate*
Methyl 4-hydroxy-3-methoxy phenyl acetate	*Methyl 4-hydroxy-3-methoxy phenyl acetate*	*Methyl 4-hydroxy-3-methoxy phenyl acetate*
Methyl 4-hydroxy-3,5-dimethoxy phenyl acetate	*Methyl 4-hydroxy-3,5-dimethoxy phenyl acetate*	*Methyl 4-hydroxy-3,5-dimethoxy phenyl acetate*
Methyl 4-hydroxy benzoate	*Methyl 4-hydroxy benzoate*	*Methyl 4-hydroxy benzoate*
Methyl 4-hydroxy-3-methoxy benzoate	*Methyl 4-hydroxy-3-methoxy benzoate*	*Methyl 4-hydroxy-3-methoxy benzoate*
Methyl 5-phenylpenta-2,4-dienoate	*Methyl 5-phenylpenta-2,4-dienoate*	*Methyl 5-phenylpenta-2,4-dienoate*
Methyl L-tyrosine	*Methyl L-tyrosine*	*Methyl L-tyrosine*
Methyl 3,4-methylene dioxy phenyl propionate	*Methyl 3,4-methylene dioxy phenyl propionate*	*Methyl 3,4-methylene dioxy phenyl propionate*
Methyl 3,4-methylene dioxy cinnamate	*Methyl 3,4-methylene dioxy cinnamate*	*Methyl 3,4-methylene dioxy cinnamate*

Common feature-based pharmacophore models

In the present section we shift our focus towards the identification of the key pharmacophoric features of substrates and the comparison of corresponding pharmacophore models for the three FAEs viz., AnFaeA, AnFaeB and TsFaeC that represent three sub-families of our proposed classification system.

First, we computed all possible pharmacophore feature mappings for the selected ligands and features using the Feature Mapping protocol in Discovery Studio 2.5 (Accelrys Software Inc, San Diego, CA, 2009, http://accelrys.com/products/discovery-studio/). The list of 15 substrates and their respective features that were used in pharmacophore modeling are given in Table 9. The feature vectors of respective substrates are given in Figure 8. The features used for mapping were hydrogen bond (HB) acceptor, HB acceptor (lipid), HB donor, hydrophobic, hydrophobic (aromatic), hydrophobic (aliphatic), positive ionizable, negative ionizable, positive charge, negative charge and ring aromatic

Table 9 Feature mapping of fifteen substrates

Substrate	Features present
Methyl cinnamate	HB Acceptor, HB Acceptor (lipid), Ring Aromatic, Hydrophobic
Methyl 2-hydroxy cinnamate	HB Acceptor, HB Acceptor (lipid), HB Donor, Ring Aromatic, Hydrophobic
Methyl 3-hydroxy cinnamate	HB Acceptor, HB Acceptor (lipid), HB Donor, Ring Aromatic, Hydrophobic
Methyl 4-hydroxy cinnamate (or) Methyl p-coumarate	HB Acceptor, HB Acceptor (lipid), HB Donor, Ring Aromatic, Hydrophobic
Methyl 3,4-dihydroxy cinnamate (or) Methyl caffeate	HB Acceptor, HB Acceptor (lipid), HB Donor, Ring Aromatic, Hydrophobic
Methyl 2-methoxy cinnamate	HB Acceptor, HB Acceptor (lipid), Ring Aromatic, Hydrophobic
Methyl 3-methoxy cinnamate	HB Acceptor, HB Acceptor (lipid), Ring Aromatic, Hydrophobic
Methyl 4-methoxy cinnamate	HB Acceptor, HB Acceptor (lipid), Ring Aromatic, Hydrophobic
Methyl 3,4-dimethoxy cinnamate	HB Acceptor, HB Acceptor (lipid), Ring Aromatic, Hydrophobic
Methyl 3,5-dimethoxy cinnamate	HB Acceptor, HB Acceptor (lipid), Ring Aromatic, Hydrophobic
Methyl 3,4,5-trimethoxy cinnamate	HB Acceptor, HB Acceptor (lipid), Ring Aromatic, Hydrophobic
Methyl 4-hydroxy-3-methoxy cinnamate (or) Methyl ferulate	HB Acceptor, HB Acceptor (lipid), HB Donor, Ring Aromatic, Hydrophobic
Methyl 3-hydroxy-4-methoxy cinnamate	HB Acceptor, HB Acceptor (lipid), HB Donor, Ring Aromatic, Hydrophobic
Methyl 4-hydroxy-3,5-dimethoxy cinnamate (or) Methyl sinapate	HB Acceptor, HB Acceptor (lipid), HB Donor, Ring Aromatic, Hydrophobic
Methyl 4-hydroxy-3-methoxy phenyl propionate	HB Acceptor, HB Acceptor (lipid), HB Donor, Ring Aromatic, Hydrophobic

Following this, we generated pharmacophores that were common to a set of active ligands. To ensure proper exploration of the ligand conformational and pharmacophoric space, the *FAST* conformation protocol (Discovey Studio 2.5) was used which employs a quasi-exhaustive systematic search to generate conformations for small molecules. Common feature pharmacophores based on the known substrates were generated for AnFaeA, AnFaeB and TsFaeC using the HipHop algorithm (Barnum et al, 1996). HipHop identifies configurations or three-dimensional spatial arrangements of chemical features common to molecules in the given training set. The algoritm evaluates training set members on basis of the type of chemical features they contain, along with the ability to adopt a conformation that allows those features to be superimposed on a particular configuration. Both active and inactive substrates of the respective enzymes were given as input and we defined that the active substrates of respective enzyme must map completely or partially to the pharmacophore; while the features from inactive substrates (on which the respective enzyme has no observed activity) must be considered as "NOT" features. This option allows broader and more diverse pharmacophores.

The resultant 10 pharmacophore models for each run were ranked based on how well the molecules mapped onto the proposed pharmacophores, as well as based on the rarity of the pharmacophore model. If a pharmacophore model is less likely to map to an inactive compound, it will be given a higher rank. As a validation, we mapped each pharmacophore model against 25 compounds, which comprise the 15 training compounds on which the pharmacophore models were built, and additional 10 compounds on which neither of AnFaeA, AnFaeB or TsFaeC can act (see Table 8). The heat map shown in Figure 9(a) indicates how well the compounds map to the respective pharmacophore models generated from the substrates of AnFaeA. The pharmacophore model 06 maps well against all of the known AnFaeA substrates; while at the same time has lower alignment scores for compounds 8 *and* 13. Both compounds are known substrates of AnFaeB and compound 13 is a known substrate of TsFaeC, but neither of the two can act on AnFaeA (Table 8). Model 06, therefore, describes with accuracy the selectivity profile of AnFaeA and it is selected as the best pharmacophore model for further analysis. Model 06 consists of three H-bond acceptors (HBA1, HBA2, and HBA3) and three hydrophobic (Hydrophobic 1, Hydrophobic 2 and Hydrophobic 3) features shown in Figure 10 (a). The pharmacophoric features of this model are perfectly mapped (with an average alignment score of 0.98) to the features of the AnFaeA active substrates.

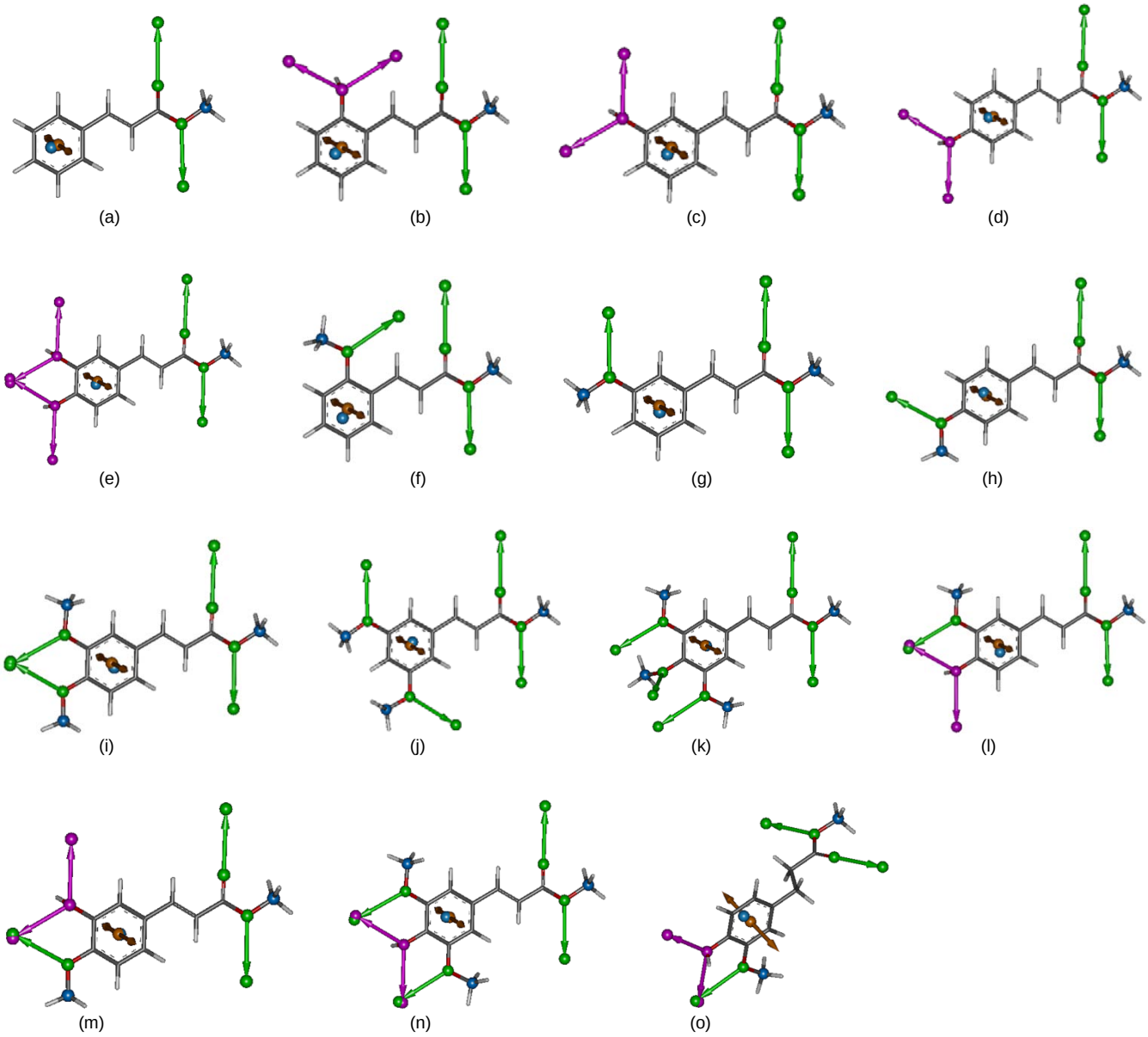


Figure 8 Pharmacophoric feature vectors of FAE substrates generated by the feature mapping protocol of Discovery studio 2.5. Colour key - Hydrogen bond donor: Magenta Vector; Hydrogen bond acceptor: Green Vector; Ring aromatic: Brown vector; Hydrophobic: blue. 3D coordinates of substrate structures were calculated with CORINA (Gasteiger et al, 1992). (a) Methyl cinnamate; (b) Methyl 2-hydroxy cinnamate; (c) Methyl 3-hydroxy cinnamate; (d) Methyl 4-hydroxy cinnamate (or) Methyl p-coumarate; (e) Methyl 3,4-dihydroxy cinnamate (or) Methyl caffeate; (f) Methyl 2-methoxy cinnamate; (g) Methyl 3-methoxy cinnamate; (h) Methyl 4-methoxy cinnamate; (i) Methyl 3,4-dimethoxy cinnamate; (j)Methyl 3,5-dimethoxy cinnamate; (k) Methyl 3,4,5-trimethoxy cinnamate; (l) Methyl 4-hydroxy-3-methoxy cinnamate (or) Methyl ferulate; (m) Methyl 3-hydroxy-4-methoxy cinnamate; (n) Methyl 4-hydroxy-3,5-dimethoxy cinnamate (or) Methyl sinapate; (o) Methyl 4-hydroxy-3-methoxy phenyl propionate.

Pharmacophore models generated for the substrates of AnFaeB and TsFaeC were also validated in a similar fashion and their heat maps are shown in Figure 9(b) and 9(c) respectively. In the case of AnFaeB, pharmacophore model 08 maps well against all AnFaeB substrates and gives very low alignment scores for compounds 2, 7, 9, 10, 11 and 14, which are substrates that selectively act on AnFaeA and/or TsFaeC (Table 8). The best performing pharmacophore model 08 for AnFaeB substrates consists of two H-bond acceptors (HBA1 & HBA2) and two hydrophobic (Hydrophobic 1& Hydrophobic 2) features shown in Figure 10(b). Finally for enzyme TsFaeC, pharmacophore model 06 aligns well with all its substrates and is able to correctly distinguish compounds 6, 8 *and* 11 as inactive. The best performing pharmacophore model 06 for TsFaeC substrates consists of two H-bond acceptors (HBA1 & HBA2) and two hydrophobic (Hydrophobic 1& Hydrophobic 2) features shown in Figure 10(c).

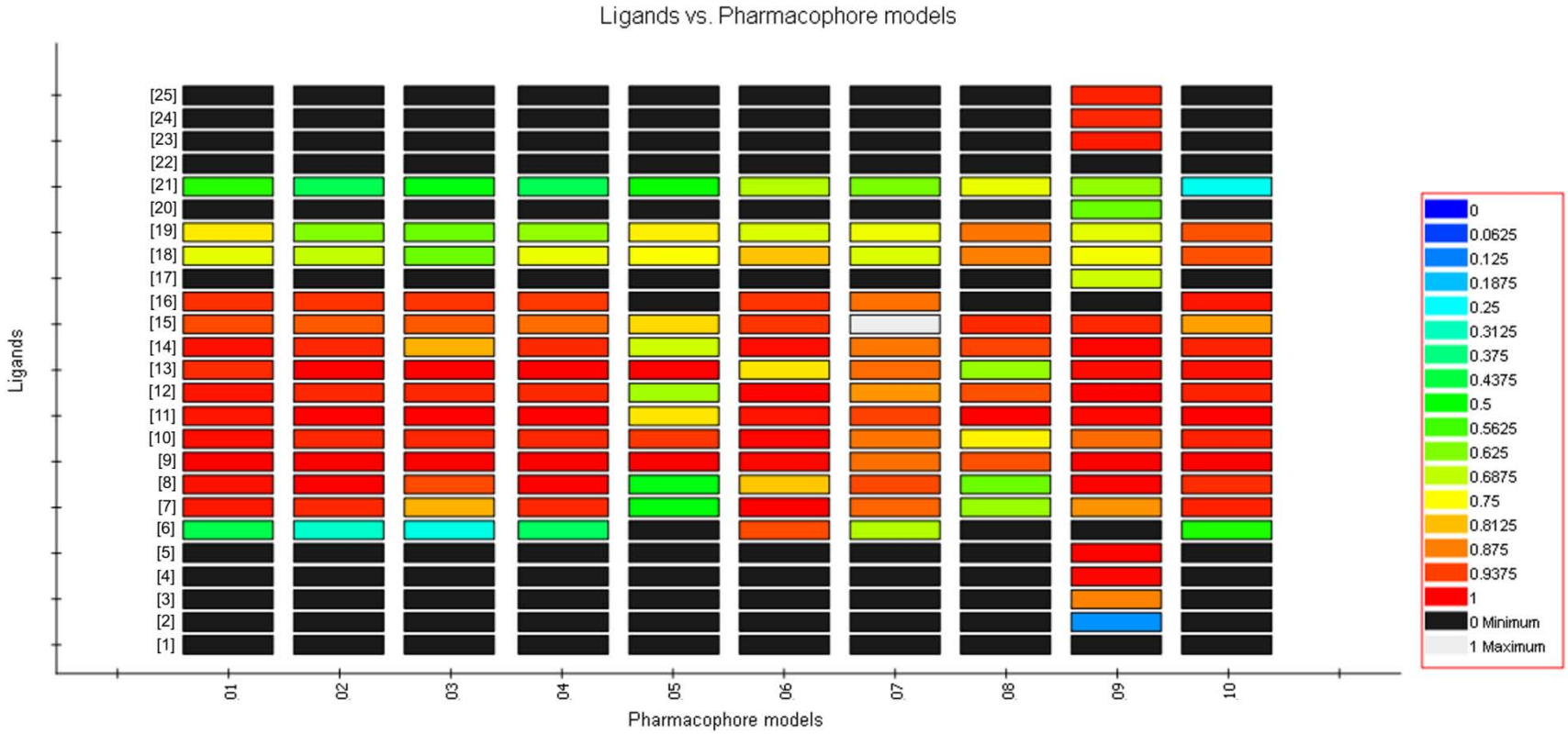


Figure 9(a) Ligand profiler heat map for pharmacophore models of AnFaeA substrates. Pharmacophore model 06 maps well against all the active substrates of AnFaeA. Substrates shown on Y-axis are [1] Methyl cinnamate; [2] Methyl 2-hydroxy cinnamate; [3] Methyl 3-hydroxy cinnamate; [4] Methyl 4-hydroxy cinnamate (or) Methyl p-coumarate; [5] Methyl 3,4-dihydroxy cinnamate (or) Methyl caffeate; [6] Methyl 2-methoxy cinnamate; **[7] Methyl 3-methoxy cinnamate;** [8] Methyl 4-methoxy cinnamate; **[9] Methyl 3,4-dimethoxy cinnamate; [10] Methyl 3,5-dimethoxy cinnamate; [11] Methyl 3,4,5-trimethoxy cinnamate; [12] Methyl 4-hydroxy-3-methoxy cinnamate (or) Methyl ferulate;** [13] Methyl 3-hydroxy-4-methoxy cinnamate; **[14] Methyl 4-hydroxy-3,5-dimethoxy cinnamate (or) Methyl sinapate; [15] Methyl 4-hydroxy-3-methoxy phenyl propionate;** [16] Methyl 3,4-dichloro phenyl propionate; [17] Methyl 4-hydroxy phenyl acetate; [18] Methyl 4-hydroxy-3-methoxy phenyl acetate; [19] Methyl 4-hydroxy-3,5-dimethoxy phenyl acetate; [20] Methyl 4-hydroxy benzoate; [21] Methyl 4-hydroxy-3-methoxy benzoate; [22] Methyl 5-phenylpenta-2,4-dienoate; [23] Methyl L-tyrosine; [24] Methyl 3,4-methylene dioxy phenyl propionate; [25] Methyl 3,4-methylene dioxy cinnamate. The heat map values show how well compounds map to pharmacophore models; higher values indicate better mapping of compounds to pharmacophore model. The colour legend corresponds with the alignment score and is in the range between 0 and 1.0 with high values above 0.9 (red) indicating good match. Substrates on which AnFaeA can act are highlighted in bold.

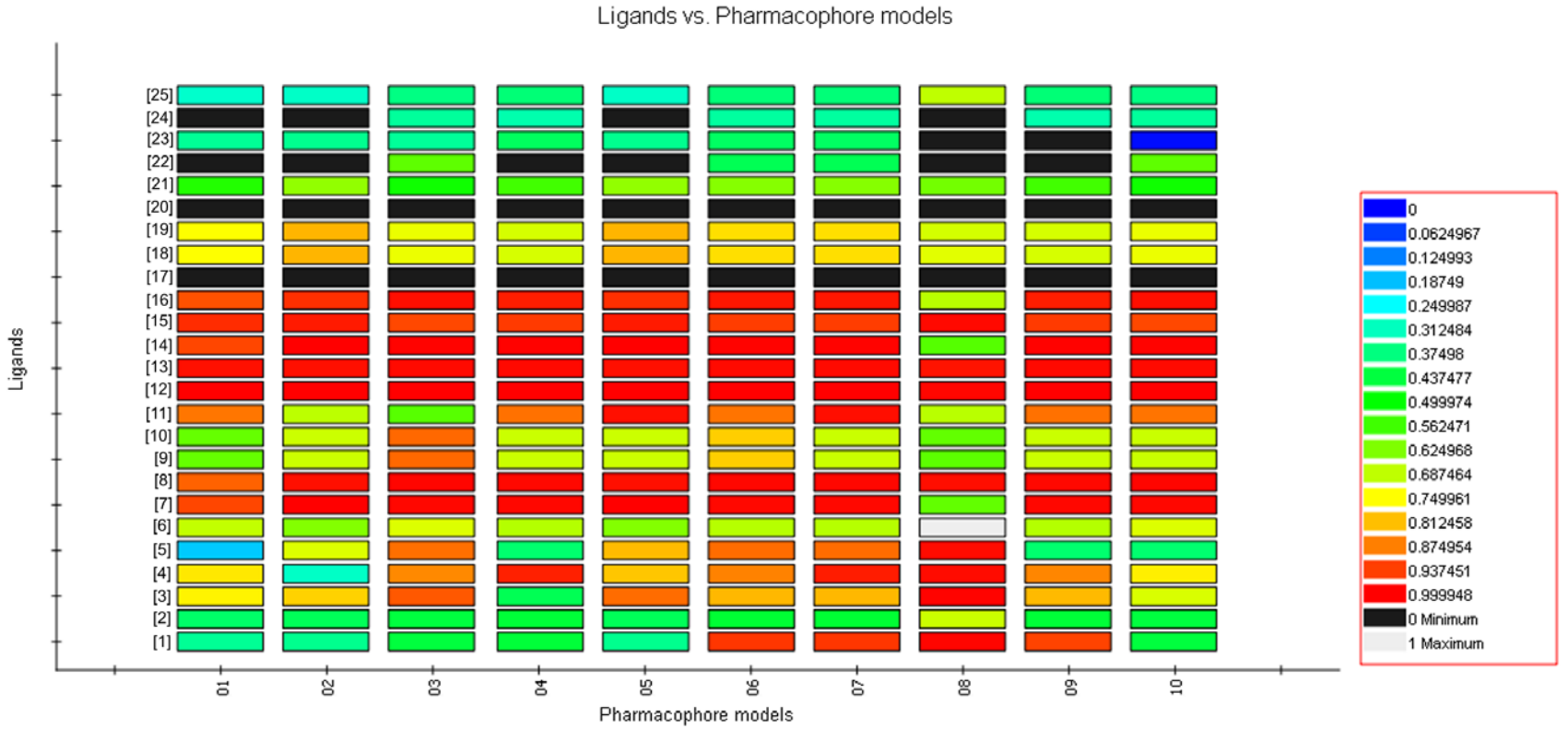


Figure 9(b) Ligand profiler heat map for pharmacophore models of AnFaeB substrates. Pharmacophore model 08 maps well against all the active substrates of AnFaeB. Substrates shown on Y-axis are **[1] Methyl cinnamate;** [2] Methyl 2-hydroxy cinnamate; **[3] Methyl 3-hydroxy cinnamate;** [**4] Methyl 4-hydroxy cinnamate (or) Methyl p-coumarate; [5] Methyl 3,4-dihydroxy cinnamate (or) Methyl caffeate; [6] Methyl 2-methoxy cinnamate;** [7] Methyl 3-methoxy cinnamate; **[8] Methyl 4-methoxy cinnamate;** [9] Methyl 3,4-dimethoxy cinnamate; [10] Methyl 3,5-dimethoxy cinnamate; [11] Methyl 3,4,5-trimethoxy cinnamate; **[12] Methyl 4-hydroxy-3-methoxy cinnamate (or) Methyl ferulate; [13] Methyl 3-hydroxy-4-methoxy cinnamate;** [14] Methyl 4-hydroxy-3,5-dimethoxy cinnamate (or) Methyl sinapate; **[15] Methyl 4-hydroxy-3-methoxy phenyl propionate;** [16] Methyl 3,4-dichloro phenyl propionate; [17] Methyl 4-hydroxy phenyl acetate; [18] Methyl 4-hydroxy-3-methoxy phenyl acetate; [19] Methyl 4-hydroxy-3,5-dimethoxy phenyl acetate; [20] Methyl 4-hydroxy benzoate; [21] Methyl 4-hydroxy-3-methoxy benzoate; [22] Methyl 5-phenylpenta-2,4-dienoate; [23] Methyl L-tyrosine; [24] Methyl 3,4-methylene dioxy phenyl propionate; [25] Methyl 3,4-methylene dioxy cinnamate. The heat map values show how well compounds map to pharmacophore models; higher values indicate better mapping of compounds to pharmacophore model. The colour legend corresponds with the alignment score and is in the range between 0 and 1.0 with high values above 0.9 (red) indicating good match. Substrates on which AnFaeB can act are highlighted in bold.

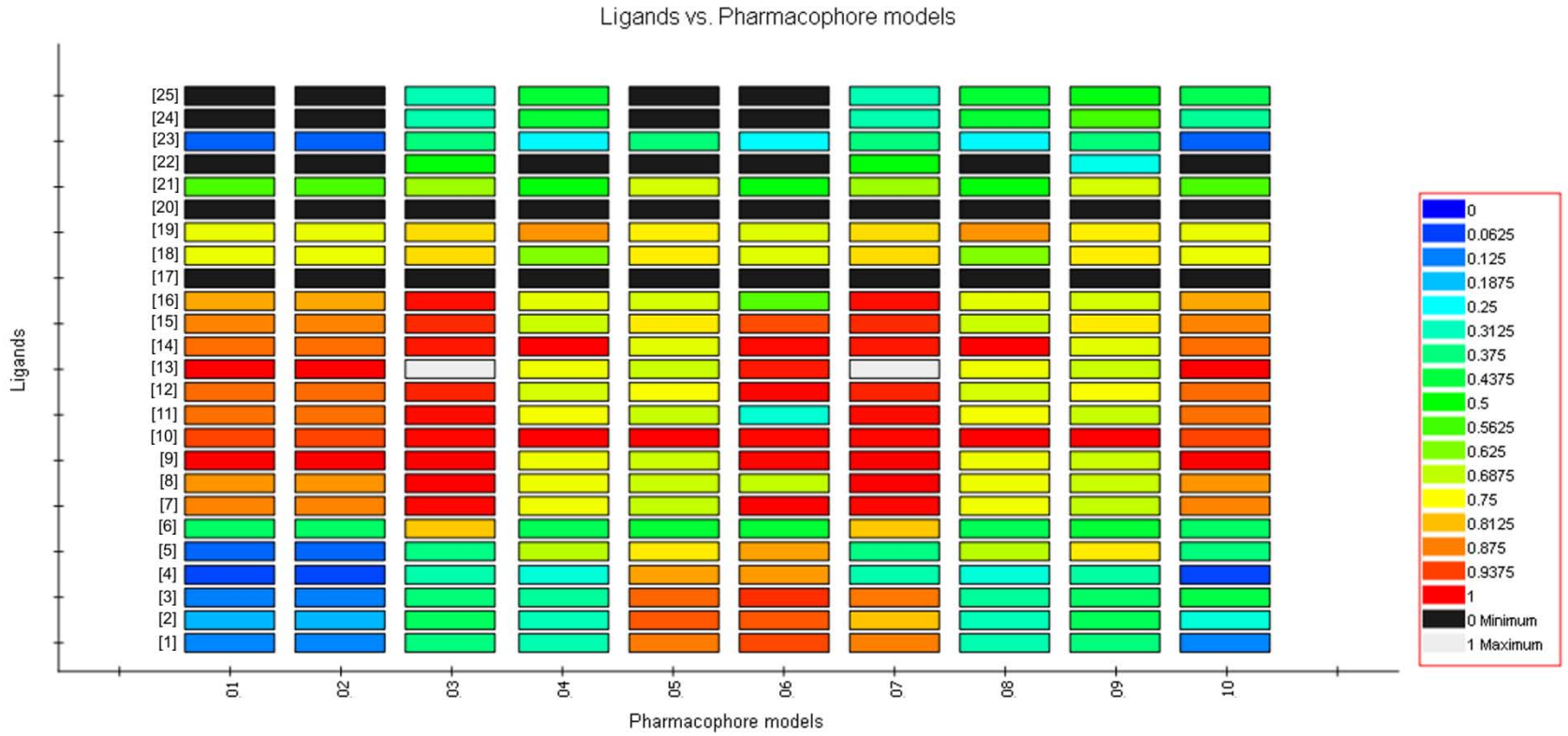
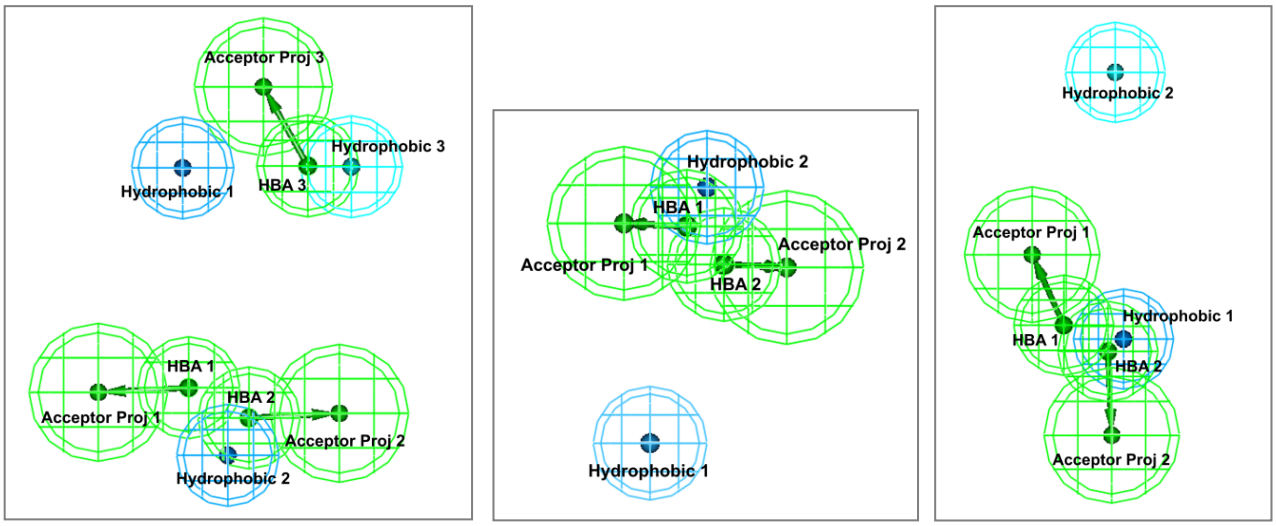


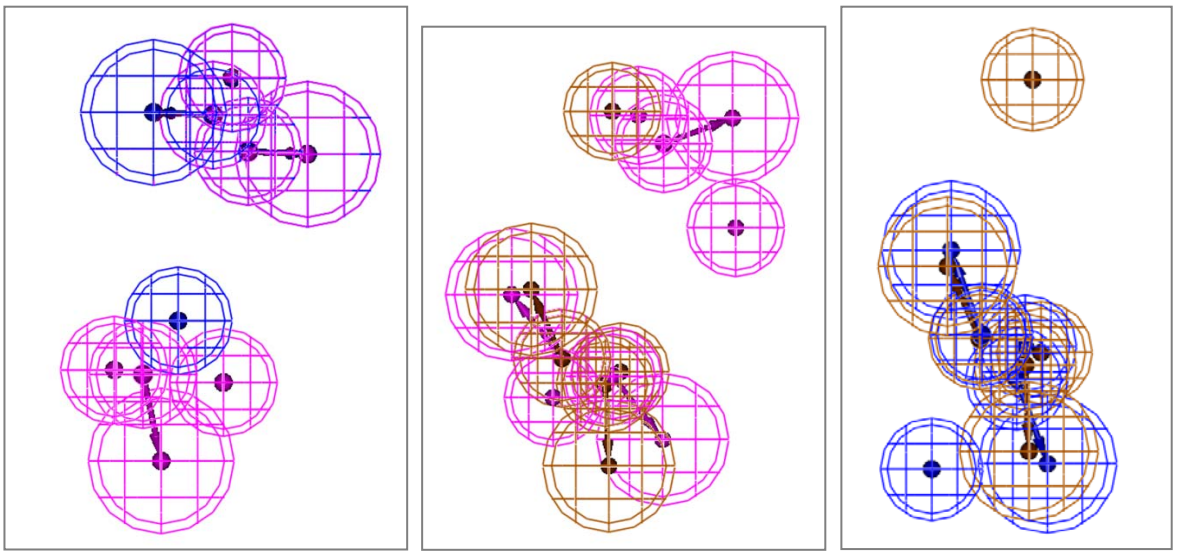
Figure 9(c) Ligand profiler heat map for pharmacophore models of TsFaeC substrates. Pharmacophore model 06 maps well against all the active substrates of TsFaeC. Substrates shown on Y-axis are **[1] Methyl cinnamate; [2] Methyl 2-hydroxy cinnamate; [3] Methyl 3-hydroxy cinnamate; [4] Methyl 4-hydroxy cinnamate (or) Methyl p-coumarate; [5] Methyl 3,4-dihydroxy cinnamate (or) Methyl caffeate;** [6] Methyl 2-methoxy cinnamate; **[7] Methyl 3-methoxy cinnamate;** [8] Methyl 4-methoxy cinnamate; **[9] Methyl 3,4-dimethoxy cinnamate; [10] Methyl 3,5-dimethoxy cinnamate;** [11] Methyl 3,4,5-trimethoxy cinnamate; **[12] Methyl 4-hydroxy-3-methoxy cinnamate (or) Methyl ferulate; [13] Methyl 3-hydroxy-4-methoxy cinnamate; [14] Methyl 4-hydroxy-3,5-dimethoxy cinnamate (or) Methyl sinapate; [15] Methyl 4-hydroxy-3-methoxy phenyl propionate;** [16] Methyl 3,4-dichloro phenyl propionate; [17] Methyl 4-hydroxy phenyl acetate; [18] Methyl 4-hydroxy-3-methoxy phenyl acetate; [19] Methyl 4-hydroxy-3,5-dimethoxy phenyl acetate; [20] Methyl 4-hydroxy benzoate; [21] Methyl 4-hydroxy-3-methoxy benzoate; [22] Methyl 5-phenylpenta-2,4-dienoate; [23] Methyl L-tyrosine; [24] Methyl 3,4-methylene dioxy phenyl propionate; [25] Methyl 3,4-methylene dioxy cinnamate. The heat map values show how well compounds map to pharmacophore models; higher values indicate better mapping of compounds to pharmacophore model. The colour legend corresponds with the alignment score and is in the range between 0 and 1.0 with high values above 0.9 (red) indicating good match. Substrates on which TsFaeC can act are highlighted in bold.



(a) (b) (c)

Figure 10 Pharmacophore models for the substrates of three enzymes viz., AnFaeA, AnFaeB and TsFaeC. (a) The best performing HipHop pharmacophore model for the substrates of AnFaeA. (b) The best performing HipHop pharmacophore model for the substrates of AnFaeB. (c) The best performing HipHop pharmacophore model for the substrates of TsFaeC. Blue: hydrophobic features; Green: H-bond acceptor features and their projections from the molecule to hydrogen-bond donors and acceptors or charged groups in the binding site.

Finally, we aligned the three pharmacophore models in order to look into their similarities and differences (see Figure 11). The differences in the pharmacophores show that a particular feature may be essential for the given enzyme, but the presence of that feature may hinder the activity of another enzyme. For example, consider the hydrophobic methoxy group features (attached to the benzene ring) of substrate [11] i.e., methyl 3,4,5-trimethoxy cinnamate. Only AnFaeA can act on methyl 3,4,5-trimethoxy cinnamate and the pharmacophore model developed for AnFaeA substrates aligns well with two of the hydrophobic methoxy group features. The pharmacophore model developed for AnFaeB substrates does not accommodate any of the hydrophobic methoxy group features, which is in line with the fact that AnFaeB cannot act on methyl 3,4,5-trimethoxy cinnamate. The pharmacophore model of TsFaeC can accommodate one of the hydrophobic methoxy features of methyl 3,4,5-trimethoxy cinnamate, and this is the reason for TsFaeC activity on methyl 3,4-dimethoxy cinnamate [9] and methyl 3,5-dimethoxy cinnamate [10] but not on methyl 3,4,5-trimethoxy cinnamate [11]. The alignments of methyl 3,4,5



(a) (b) (c)

Figure 11 Comparison of pharmacophore models for the substrates of three enzymes viz., AnFaeA, AnFaeB and TsFaeC. (a) Alignment of AnFaeA and AnFaeB pharmacophore models. (b) Alignment of AnFaeA and TsFaeC pharmacophore models. (c) Alignment of AnFaeB and TsFaeC pharmacophore models. For better view of alignments, Pharmacophore models for AnFaeA, AnFaeB and TsFaeC were colored in Magenta, Blue and Brown respectively. For details of respective pharmacophoric features see Fig 10.

The alignments of methyl 3,4,5-trimethoxy cinnamate [11] with respective pharmacophores are depicted in Figure 12(a), (b) and (c). Hydrophobic methoxy group features are depicted as polyhedrons in Figure 12(d), (e) and (f) which shows that presence of more than one hydrophobic methoxy groups attached to the benzene ring of the substrates makes AnFaeB incapable for acting on them. Whereas, TsFaeC can accommodate one hydrophobic methoxy group attached to the benzene ring of the substrate, but presence of more than two hydrophobic methoxy groups makes it incapable to act on the substrate.

The pharmacophore models developed from AnFaeA, AnFaeB and TsFaeC substrates represent all members of the sub-families 12A, 4A and 4B respectively (see Table 7) and can, thus, be used for the prediction of their substrate binding profiles. Availability of sequence and binding activity data for FAEs of other sub-families will make possible the development of pharmacophore models that will represent the members of their respective sub-families, as we propose them in the present work.

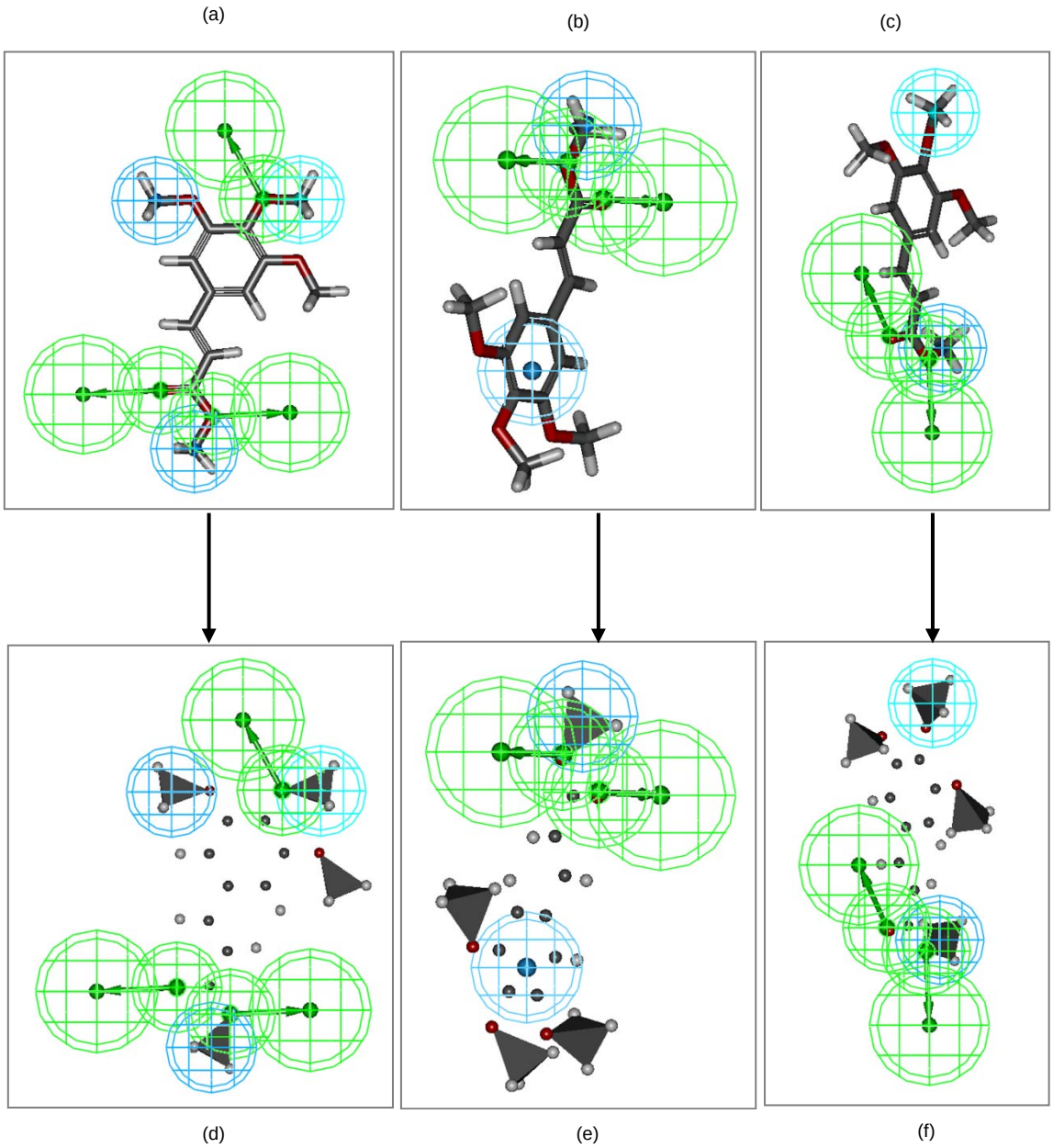


Figure 12 Comparison of hydrophobic features in the pharmacophore models developed for the substrates of AnFaeA, AnFaeB and TsFaeC. (a) Alignment of methyl 3,4,5-trimethoxy cinnamate with the pharmacophore model of AnFaeA substrates. (b) Alignment of methyl 3,4,5-trimethoxy cinnamate with the pharmacophore model of AnFaeB substrates. (c) Alignment of methyl 3,4,5-trimethoxy cinnamate with the pharmacophore model of TsFaeC substrates. (d), (e), (f): The hydrophobic methoxy group features depicted in polyhedrons. For details of respective pharmacophoric features see Fig 10.

Conclusions and future prospects

While the knowledge on production and characterization of FAEs has been increasing at a rapid and exciting rate, publication of primary sequences of characterized FAEs and structure- function relationship data are still at low pace. The recent growth in reports showing FAEs with unlimited properties and applications pushed forward to a new classification platform considering FAEs from Bacteria, Fungi and Plantae. Even though FAEs possess common characteristics, like the classic constellation of the Ser-His-Asp triad, variations in amino acid sequences forming surface loops and additional domains allow them to accommodate diverse substrates. By using the properties of the whole sequence we propose a new classification system for FAEs resulting into 12 distinct families, while by careful inspection of the catalytic residues constellation in the sequences of each family we were able to further divide FAEs into more informative sub-families. FAEs empirically characterized as type A-D by Crepin et al (2004), are correctly placed in different sub-families in our classification scheme. At the same time, FAEs of the same type are primarily found in the same sub-family, with the exception of type B FAEs that are shared among three sub-families, demonstrating the biological relevance of our method. We should emphasize the fact that the classification system that we propose does not contradict but rather significantly expands the current knowledge in the area and allows a systematic understanding of the mode of action of FAEs.

In addition, the development of pharmacophore models for specific FAE sub-families will have a huge impact on the application of members of the particular group to completely novel and unexpected substrates. Virtual screening with the developed pharmacophores of chemical and natural compound databases could reveal unique opportunities for FAEs-based-biocatalytic modifications to synthesize compounds with altered or improved medicinal properties. Acquisition of biochemical data for FAEs that belong to each of the proposed families will further complete our descriptor-based classification system. We are confident that it will also provide researchers and industries with the toolbox from which to select FAEs for suitable reactions and applications.

Acknowledgements

The authors thank Tien Luu (Lead Scientific Specialist, Life Science, Accelrys Software Inc) for discussions in pharmacophore generation and support through providing Discovery Studio 2.5 Software package (2009). The authors are grateful to Olivier Taboureau for fruitful discussions. Funding support for this research was provided by The Swedish Research Council (Vetenskapsrådet). DGU thank Center for Biological Sequence Analysis at DTU for providing computing resources. IK and GPA acknowledge financial support from the Danish Research Council for Technology and Production Sciences. The authors have declared no conflict of interest.

Appendix A. Supplementary data

The following Supplementary Files associated with this article can be obtained though e-mail from D.B.R.K. Gupta Udatha (gupta.udatha@chalmers.se)

Supplementary File 1:

Table S1: FAEs from fungi, bacteria and plantae kingdoms along with their reported hydrolytic substrate specificities.
Table S2: FAEs and reported synthetic applications.
Table S3: FAE-related sequences from fungi, sorted according to Taxonomy lineage.
Table S4: FAE-related sequences from bacteria, sorted according to Taxonomy lineage.
Table S5: FAE-related sequences from plantae, sorted according to Taxonomy lineage.
Table S6: FAE-related sequences from protists, sorted according to Taxonomy lineage.
Table S7: FAE-related sequences – respective lengths of signal peptide and mature sequence.
Table S8: Distribution of FAE-related sequences from all kingdoms among the bootstrapped rectangular phylogram. Compared with the Taxonomy lineage shown in Tables S1 to S4, the lineage presented here is jumbled according to bootstrapped phylogram. Distribution of seventeen characterized FAE and their type is also indicated according to the classification system of Crepin et al (2004).

Supplementary File 2:

Table S9: Thirteen FAE clusters of descriptor set DS14; protein name, organism and kingdom of respective members are given. FAE type of experimentally characterized sequences is designated in the last column according to the classification system of Crepin et al (2004).

Supplementary File 3:

Table S10: Superfamilies assigned to FAE and putative FAE sequences by UFO server.
Table S11: Blocks graphical map positions in sequences of respective clusters. Blocks graphical map showing conserved domains in the members of respective clusters.
Table S12: Nucleophilic elbow [GXSXG] positions in FAE and putative FAE sequences detected using BioEdit program.
Table S13: Catalytic triad residues predicted by INTREPID server. The catalytic triad residues with high INTREPID information-theoretic score were selected from Top 25 functional residues output.

Supplementary File 4:

Table S14: Classification system of FAEs covering fungal, bacterial and plantae kingdoms. Distribution of known types of FAEs (Crepin et al, 2004) among the sub-families is also indicated.

Akin DE, Borneman WS, Rigsby LL, Martin SA. p-Coumaroyl and feruloyl arabinoxylans from plant cell walls as substrates for ruminal bacteria. Appl Environ Microbiol 1993; (2): 644–647

Akin DE. Plant cell wall aromatics: influence on degradation of biomass. Biofuels Bioprod Bioref 2008; 2:288-303

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25(17):3389-402

Andreasen MF, Kroon P, Williamson G, Garcia-Conesa MT. Intestinal release and uptake of phenolic antioxidant diferulic acids. Free Radical Biol Med 2001; 31:304-314

Andersen A, Svendsen A, Vind J, Lassen SF, Hjort C, Borch K, Patkar SA. Studies on ferulic acid esterase activity in fungal lipases and cutinases. Colloids Surf B Biointerfaces 2002; 26:47–55

Asther M, Haon M, Roussos S, Record E, Delattre M, et al. Feruloyl esterase from Aspergillus niger a comparison of the production in solid state and submerged fermentation. Process Biochem 2002; 38: 685- 691

Aurilia V, Parracino A, Saviano M, Rossi M, D'Auria S. The psychrophilic bacterium Pseudoalteromonas halosplanktis TAC125 possesses a gene coding for a cold-adapted feruloyl esterase activity that shares homology with esterase enzymes from γ-proteobacteria and yeast. Gene 2007; 397:51–57

Barnum D, Greene J, Smellie A, Sprague P. Identification of common functional configurations among molecules. J Chem Inf Comput Sci 1996; 36(3):563-71

Bartolomé B, Faulds CB, Tuohy M, Hazlewood GP, Gilbert HJ, Williamson G. Influence of different xylanases on the activity of ferulic acid esterase on wheat bran. Biotechnol Appl Biochem 1995; 22:65-73

Bartolomé B, Faulds CB, Kroon PA, Waldron K, Gilbert HJ, Hazlewood G, Williamson G. An *Aspergillus niger* Esterase (Ferulic Acid Esterase III) and a Recombinant *Pseudomonas fluorescens* subsp. *Cellulose* Esterase (XylD) Release a 5-59 Ferulic Dehydrodimer (Diferulic Acid) from Barley and Wheat Cell Walls. Appl Environ Microbiol 1997; 63(1):208-212

Bartolomé B, Gómez-Cordovés C, Sancho AI, Díez N, Ferreira P, Soliveri J, Copa-Patiño JL. Growth and release of hydroxycinnamic acids from Brewer's spent grain by *Streptomyces avermitilis* CECT 3339. Enzyme Microb Technol 2003; 32:140–144

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 2004; 340:783-795

Benner SA, Chamberlin SG, Liberles DA, Govindarajan S, Knecht L. Functional inferences from reconstructed evolutionary biology involving rectified databases-an evolutionarily grounded approach to functional genomics. Res Microbiol 2000; 151:97-106

Benoit I, Navarro D, Marnet N, Rakotomanomana N, et al. Feruloyl esterases as a tool for the release of phenolic compounds from agro-industrial by-products. Carbohydr Res 2006; 341:1820–1827

Benoit I, Asther M, Sulzenbacher G, Record E, Marmuse L, Parsiegla G, Gimbert I, Asther M, Bignon C. Respective importance of protein folding and glycosylation in the thermal stability of recombinant feruloyl esterase A. FEBS Lett 2006; 580(25):5815-5821

Benoit I, Danchin EGJ, Bleichrodt RJ, De Vries RP. Biotechnological applications and potential of fungal feruloyl esterases based on prevalence, classification and biochemical diversity. Biotechnol Lett 2008; 30:387–396

Bhasin M, Raghava GPS. Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition. J Bio Chem 2004; 279:23262-23266

Bhasin M, Garg A, Raghava GP: PSLpred: prediction of subcellular localization of bacterial proteins. Bioinformatics 2005, 21(10):2522-2524.

Bhaskaran R, Ponnuswamy PK. Positional flexibilities of amino acid residues in global proteins. Int J Peptide Protein Res 1988; 32:241-235

Bigelow CC. On the average hydrophobicity of proteins and the relation between it and protein structure. J Theor Biol 1967; 16(2):187-211

Blum DL, Kateva IA, Li XL, Ljungdahl LG. Feruloyl Esterase Activity of the *Clostridium thermocellum* Cellulosome Can Be Attributed to Previously Unknown Domains of XynY and XynZ. J Bacteriol 2000; 182(5):1346–1351

Borneman WS, Hartley RD, Morrison WH, Akin DE, Ljungdahl LG. Feruloyl and p-coumaroyl esterase from anaerobic fungi in relation to plant cell wall degradation. Appl Microbiol Biotechnol 1990; 33:345-351.

Borneman WS, Ljungdahl LG, Hartley RD, Akin DE. Purification and partial characterization of two feruloyl esterases from the anaerobic fungus *Neocallimastix* strain MC-2. Appl Environ Microb 1992; 57:3762-3766

Bouzid O, Navarro D, Roche M, Asther M, Haon M, et al. Fungal enzymes as a powerful tool to release simple phenolic compounds from olive oil by-product. Process Biochem 2005; 40:1855-1862

Brézillon C, Kroon PA, Faulds CB, Brett GM, Williamson G. Novel ferulic acid esterases are induced by growth of Aspergillus niger on sugar-beet pulp. Appl Microbiol Biotechnol 1996; 45:371-376

Broto P, Moreau G, Vandicke C. Molecular structures: perception, autocorrelation descriptor and SAR studies. Eur J Med Chem 1984; 19:71-78

Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 2003; 31:3692-3697

Cai CZ, Wang WL, sun LZ, Chen YZ. Protein function classification via support vector machine approach. Math Biosci 2003; 185(2):111-122

Cai CZ, Han LY, Ji ZL, Chen YZ. Enzyme family classification by support vector machines. Proteins 2004; 55(1):66-76

Castanares A, Wood TM. Purification and characterization of a feruloyl/p-coumaroyl esterase from solid-state cultures of the aerobic fungus Penicillium pinophilum. Biochem Soc Trans 1992; 20(3):275S

Chang CC, Lin CJ, LIBSVM : a library for support vector machines, 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm

Charton M. Protein folding and the genetic code: An alterative quantitative model. J Theor Biol 1981; 91:115-123

Charton M, Charton BI. The Structure Dependence of Amino Acid Hydrophobicity parameters. J Theor Biol 1982; 99:629-644

Chothia C. The Nature of the Accessible and buried surface in proteins. J Mol Biol 1976; 105:1-14

Chou KC. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 2000; 278:477-483

Chou KC. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Structure Function and Genetics 2001; 43:246-255

Chou KC, Cai YD. Prediction of protein subcellular locations by GO-FunD-PseAA predictor. Biochem Biophys Res Commun 2004; 320(4):1236-9.

Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 2005; 21:10-19

Cid H, Bunster M, Canales M, Gazitúa F. Hydrophobicity and structural classes in proteins. Protein Eng 1992; 5(5):373-5

Crepin VF, Faulds CB, Connerton IF. Production and characterization of the Talaromyces stipitatus feruloyl esterase FAEC in Pichia pastoris: identification of the nucleophilic serine. Protein Expr Purif 2003; 29:176–184

Crepin VF, Faulds CB, Connerton IF. A non-modular type B feruloyl esterase from Neurospora crassa exhibits concentration-dependent substrate inhibition. Biochem J 2003; 370:417-427

Crepin VF, Faulds CB, Connerton IF. Functional classification of the microbial feruloyl esterases. Appl Microbiol Biotechnol 2004; 63:647–652

Dalrymple BP, Swadling Y, Cybinski DH, Xue GP. Cloning of a gene encoding cinnamoyl ester hydrolase from the ruminal bacterium Butyrivibrio fibrisolvens E14 by a novel method. FEMS Microbiol Lett 1996; 143(2-3):115-120.

Dayhoff MO,Schwartz RM, Orcutt BC. Altas of Protein Sequence and Structure. National Biomedicinal Research Foundation, Washington D.C.1979; 5(3):345-352

Dinisa MJ, Bezerrab RMF, Nunesc F, Diasb AA, Guedesa CV, et al. Modification of wheat straw lignin by solid state fermentation with white-rot fungi. Biores Technol 2009; 100:4829-4835

Discovery Studio Client Version 2.5; Accelrys Software Inc, San Diego, CA, 2009.

Dodd D, Kocherginskaya SA, Spies MA, Beery KE, Abbas CA, Mackie RI, Cann IKO. Biochemical Analysis of a β-D-Xylosidase and a Bifunctional Xylanase-Ferulic Acid Esterase from a Xylanolytic Gene Cluster in *Prevotella ruminicola* 23. J Bacteriol 2009; 191(10): 3328-3338

Donaghy J, McKay AM. Purification and characterization of a feruloyl esterase from the fungus *Penicillium expansum*. J Appl Microbiol 1997; 83:718-726

Donaghy JA, McKay AM. Measurement of feruloyl/p-coumaroyl esterase by capillary zone electrophoresis. World J Microbiol Biotechnol 1995; 11:160-162

Donaghy JA, McKay AM. Production of feruloyl/rho-coumaroyl esterase activity by Penicillium expansum, Penicillium brevicompactum and Aspergillus niger. J Appl Bacteriol 1995; 79(6):657-662

Donaghy JA, Bronnenmeier K, Soto-Kelly PF, McKay AM. Purification and characterization of an extracellular feruloyl esterase from the thermophilic anaerobe *Clostridium stercorarium.* J Appl Microbiol 2000; 88:458–466

Dubchak I, Muchnick I, Holbrook SR, Kim SH. Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci USA 1995, 92:8700-8704.

Dubchak I, Muchnick I, Mayor C, Dralyuk I, Kim SH. Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. Proteins 1999; 35:401-407

Dudoit S, Fridlyand J, Speed TP. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Technical report no. 576, Department of Statistics, University of California, 2000.

Dysvik B, Jonassen I. J-Express: exploring gene expression data using Java. Bioinformatics 2001;17(4):369-70

Ewing TJA, Kuntz ID. Critical evaluation of search algorithms for automated molecular docking and database screening. Journal of Computational Chemistry 1998; 18(9):1175-1189.

Faulds CB, Williamson G. The purification and characterization of 4-hydroxy-3-methoxycinnamic (ferulic) acid esterase from *Streptomyces olitrochromogenes.* J Gen Microbiol 1991; 137:2339-2345

Faulds CB, Williamson G. Release of Ferulic acid from plant polysaccharides by ferulic acid esterase from *Streptomyces olivochromogenes*, Carbohydr Polym 1993; 21:153-155

Faulds CB, Williamson G. Ferulic acid esterase from Aspergillus niger: purification and partial characterization of two forms from a commercial source of pectinase. Biotechnol Appl Biochem 1993; 17(3):349–359

Faulds CB, Williamson G. Purification and characterization of a ferulic acid esterase (FAE-III) from *Aspergillus niger*: specificity of the phenolic moiety and binding to microcrystalline cellulose. Microbiology 1994; 140:779-787

Faulds CB, Ralet MC, Williamson G, Hazlewood GP, Gilbert HJ. Specificity of an esterase (XYLD) from *Pseudomonas fluorescens* subsp. *cellulosa*. Biochim Biophys Acta 1995; 1243:265–269

Faulds CB, Kroon PA, Saulnier L, Thibault JF, Williamson G. Release of ferulic acid from maize bran and derived oligosaccharides by *Aspergillus niger* esterases. Carbohydr Polym 1995; 27:187-190

Faulds CB, Zanichelli D, Crepin VF, Connerton IF, Juge N, Bhat MK, Waldron KW. Specificity of feruloyl esterases for water-extractableand water-unextractable feruloylated polysaccharides: influence of Xylanase. J Cereal Sci 2003; 38:281-288

Faulds CB, Molina R, Gonzalez R, Husband F, Juge N, Sanz-Aparicio J, Hermoso JA. Probing the determinants of substrate specificity of a feruloyl esterase, AnFaeA, from Aspergillus niger. FEBS J 2005; 272(17):4362-4371

Fazary AE, Ismadji S, Ju YH. Stability and solubility studies of native and activated Aspergillus awamori feruloyl esterase. J Mol Catal B Enzym 2009; 59:190-196

Feng ZP, Zhang CT. Prediction of membrane protein types based on the hydrophobic index of amino acids. J Protein Chem 2000; 19:269-275

Ferguson LR, Harris PJ. Protection against cancer by wheat bran: role of dietary fibre and phytochemicals. Eur J Cancer Prev 1999; 8:17-25

Fillingham IJ, Kroon PA, Williamson G, Gilbert HJ, Hazlewood GP. A modular cinnamoyl ester hydrolase from the anaerobic fungus Piromyces equi acts synergistically with xylanase and is part of a multiprotein cellulose-binding cellulase–hemicellulase complex. Biochem J 1999; *343*:215-224

Finn R, Tate J, Mistry J, Coggill P, Sammut J, Hotz H, Ceric G, Forslund K, Eddy S, Sonnhammer E, Bateman A. The Pfam protein family database. Nucleic Acids Res 2008; 36:281–288

Gao QB, Wang ZZ. Classification of G-protein coupled receptors at four levels. Protein Eng Des Sel 2006; 19(11):511-516

Garcia-Conesa MT, Kroon PA, Ralph J, Mellon FA, Colquhoun IJ, Saulnier L, Thibault JF, Williamson G. A cinnamoyl esterase from Aspergillus niger can break plant cell wall cross-links without release of free diferulic acids. Eur J Biochem 1999; 266:644-652

García-Conesa MT, Østergaard P, Kauppinen S, Williamson G. Hydrolysis of diethyl diferulates by a tannase from *Aspergillus oryzae.* Carbohydrate Polymers 2001; 44:319–324

Garcia-Conesa MT, Crepin VF, Goldson AJ, Williamson G, Cummings NJ, Connerton IF, Faulds CB, Kroon PA. The feruloyl esterase system of *Talaromyces stipitatus*: production of three discrete feruloyl esterases, including a novel enzyme, TsFaeC, with a broad substrate specificity. J Biotechnol 2004; 108:227–241

Garrigues GE, Cho DR, Rubash HE, Goldring SR, Herndon JH, Shanbhag AS.Gene expression clustering using self-organizing maps: analysis of the macrophage response to particulate biomaterials. Biomaterials 2005; 26(16):2933-45

Gasteiger J and Engel T. Chemoinformatics: A Textbook. Weinheim, WILEY-VCH, 2003

Gasteiger J, Rudolph C, Sadowski J. Automatic generation of 3D-atomic coordinates for organic molecules. Tetrahedron Comp Method 1992; 3: 537–547

Giuliani S, Piana1 C, Setti L, Hochkoeppler A, Pifferi PG, Williamson G, Faulds CB. Synthesis of pentylferulate by a feruloyl esterase from *Aspergillus niger* using water-in-oil microemulsions. Biotechnol Lett 2001; 23:325–330

Goldstone DC, Villas-Bôas SG, Till M, Kelly WJ, Attwood GT, Arcus VL. Structural and functional characterization of a promiscuous feruloyl esterase (Est1E) from the rumen bacterium *Butyrivibrio proteoclasticus*. Proteins 2010; 78(6):1457-69

Grantham R. Amino acid difference formula to help explain protein evolution. Science 1974; 185:862-864

Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 1999; 41:95–98

Han LY, Cai CZ,Ji ZL, Cao ZW, Cui J, Chen YZ. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. Nucleic Acids Res 2004; 32(21):6437–6444

Hatzakis NS, Daphnomili D, Smonou I. Ferulic acid esterase from Humicola Insolens catalyzes enantioselective transesterification of secondary alcohols. Journal of Molecular Catalysis B: Enzymatic 2003; 21(4-6):309-311

Hegde S, Kavitha S, Varadaraj MC, Muralikrishna G. Degradation of cereal bran polysaccharide-phenolic acid complexes by Aspergillus niger CFR 1105. Food chem 2006; 96:14-19

Henikoff S, Henikoff JG, Alford WJ, Pietrokovski S. Automated construction and graphical presentation of protein blocks from unaligned sequences. Gene 1995;163(2):GC17-GC26

Hermoso JA, Sanz-Aparicio J, Molina R, Juge N, Gonzalez R, Faulds CB. The crystal structure of feruloyl esterase A from Aspergillus niger suggests evolutive functional convergence in feruloyl esterase family. J Mol Biol 2004; 338:495–506

Horne DS. Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. Biopolymers 1988; 27:451-477

Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification, 2009. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/gui

Humberstone FJ, Briggs DE. Partial purification of ferulic acid esterase from malted barley. J Inst Brew 2002; 108(4):439-443

Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Regula R. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 2007; 8:460

Kaiser D, Terfloth L, Kopp S, Schulz J, de Laet R, Chiba P, Ecker GF, Gasteiger J. Self-organizing maps for identification of new inhibitors of P-glycoprotein. J Med Chem 2007; 50(7):1698-702

Kanauchi M, Watanabe S, Tsukada T, Atta K, Kakuta T, Koizumi T. Purification and Characteristics of Feruloyl Esterase from *Aspergillus awamori* G-2 Strain. J Food Sci 2008; 73(6):C458-463

Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines. Bioinformatics 2002; 18:147-159

Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK. Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation 2001; 13(3):637-649

Kheder F, Delaunay S, Abo-Chameh G, Paris C, Muniglia L, Girardin M. Production and biochemical characterization of a type B ferulic acid esterase from *Streptomyces ambofaciens*. Can J Microbiol 2009; 55(6): 729-738

Kikuzaki H, Hisamoto M, Hirose K, Akiyama K, Taniguchi H. Antioxidant properties of ferulic acid and its related compounds. J Agric Food Chem 2002; 50(7):2161-2168.

Kim H, Park H. Protein secondary structure prediction based on an improved support vector machines approach. Protein Eng 2003; 16(8):553-60

Knoshaug EP, Selig MJ, Baker JO, Decker SR, Himmel ME, Adney WS. Heterologous Expression of Two Ferulic Acid Esterases from Penicillium funiculosum. Appl Biochem Biotechnol 2008; 146:79-87

Kohavi R, A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI 1995; 2(12):1137–1143

Kohonen T. Self-Organizing Maps. Third extended edition. Springer, 2001

Koseki T, Furuse S, Iwano K, Matsuzawa H. Purification and Characterization of Feruloyl esterase from Aspergillus awamori. Biosci Biotechnol Biochem 1998; 62(10):2032-2034

Koseki T, Takahashi K, Fushinobu S, Iefuji H, Iwano K, Hashizume K, Matsuzawa H. Mutational analysis of a feruloyl esterase from Aspergillus awamori involved in substrate discrimination and pH dependence. Biochim Biophys Acta 2005; 1722: 200– 208

Koseki T, Takahashi K, Handa T, Yamane Y, Fushinobu S, Hashizume K. N-Linked oligosaccharides of Aspergillus awamori Feruloyl esterase are important for thermostability and catalysis. Biosci Biotechnol Biochem 2006; 70(10):2476-2480

Koseki T, Hori A, Seki S, Murayama T, Shiono Y. Characterization of two distinct feruloyl esterases, AoFaeB and AoFaeC, from Aspergillus oryzae. Appl Microbiol Biotechnol 2009; 83:689-696

Koseki T, Fushinobu S, Ardiansyah, Shirakawa H, Komai M. Occurrence, properties, and applications of feruloyl esterases. Appl Microbiol Biotechnol 2009; 84(5):803-10

Kroon PA, Faulds CB, Brezillon C, Williamson G. Methyl phenylalkanoates as substrates to probe the active sites of esterases, Eur J Biochem 1997; 248:245-251

Kroon PA, Williamson G, Fish NM, Archer DB, Belshaw NJ. A modular esterase from Penicillium funiculosum which releases ferulic acid from plant cell walls and binds crystalline cellulose contains a carbohydrate binding module. Eur J Biochem 2000; 267(23):6740-52

Kumar M, Thakur V, Raghava GP. COPid: composition based protein identification. In Silico Biol 2008; 8(2):121-128

Lao DM, Arai M, Ikeda M, Shimizu T. The presence of signal peptide significantly affects transmembrane topology prediction. Bioinformatics 2002; 18:1562-1566

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. ClustalW and ClustalX version 2.0. Bioinformatics 2007; 23(21):2947-2948.

Laszlo JA, Compton DL, Li XL. Feruloyl esterase hydrolysis and recovery of ferulic acid from jojoba meal. Ind Crop Prod 2006; 23:46-53

Latha GM, Srinivas P, Muralikrishna G. Purification and Characterization of Ferulic Acid Esterase from Malted Finger Millet (*Eleusine coracana*, Indaf-15). J Agric Food Chem 2007; 55:9704-9712

Lee JA, Verleysen M.Self-organizing maps with recursive neighborhood adaptation. Neural Netw 2002; 15(8-9):993-1003

Leeuwen MJFS, Vincken JP, Schipper D, Voragen AGJ, Beldman G. Acetyl esterases of *Aspergillus niger*: purification and mode of action on pectins, Prog Biotechnol 1996; 793–798

Lesage-Meessen L, Delattre M, Haon M, Thibault JF, Ceccaldi BC, Brunerie P, Asther M. A two-step bioconversion process for vanillin production from ferulic acid combining Aspergillus niger and Pycnoporus cinnabarinus. J Biotechnol 1996; 50(2-3):107-13

Levasseur A, Navarro D, Punt PJ, Belaıch JP, Asther M, Record E. Construction of engineered bifunctional enzymes and their overproduction in *Aspergillus niger* for improved enzymatic tools to degrade agricultural by-products, Appl Environ Microbiol 2005; 71:8132-8140

Levasseur A, Gouret P, Lesage-Meessen L, Michèle Asther, Marcel Asther, Record E, Pontarotti P. Tracking the connection between evolutionary and functional shifts using fungal lipase/feruloyl esterase A family. BMC Evol Biol 2006; 6:92

Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ. PROFEAT: A Web Server for Computing Structural and Physicochemical Features of Proteins and Peptides from Amino Acid Sequence. Nucleic Acids Res.Jul 1, 2006; 34(Web Server issue):W32-7

Lima T, Auchincloss AH, Coudert E, Keller G, Michoud K, Rivoire C, Bulliard V, de Castro E, Lachaize C, Baratin D, Phan I, Bougueleret L, Bairoch A. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. Nucleic Acids Res 2009; 37:D471–D478

Lin Z, Pan XM. Accurate prediction of protein secondary structural content. J Protein Chem 2001; 20:217-220

Lin HH, Han LY, Zhang HL, Zheng CJ, Xie B, Chen YZ. Prediction of the Functional Class of Lipid-Binding Proteins from Sequence Derived Properties Irrespective of Sequence Similarity. J Lipid Res 2006; 47(4):824-31

MacKay, David. Information Theory, Inference and Learning Algorithms. Cambridge University Press, 2003

Mandalari G, Bisignano G, Lo Curto RB, Waldron KW, Faulds CB. Production of feruloyl esterases and xylanases by *Talaromyces stipitatus* and *Humicola grisea* var. *thermoidea* on industrial food processing by-products. Bioresour Technol 2008; 99:5130-5133

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein–protein interactions from genome sequences. Science 1999; 285:751–753

McAuley KE, Svendsen A, Patkar SA, Wilson KS. Structure of a feruloyl esterase from Aspergillus niger. Acta Crystallogr D Biol Crystallogr 2004; 60(5):878-87

McCallum JA, Taylor IE, Toers GN. Spectrophotometric assay and electrophoretic detection of *trans*-feruloyl esterase activity. Anal Biochem 1991; 196: 360−366

McCrae SI, Leith KM, Gordon AH, Wood TM. Xylan-degrading enzyme system produced by the fungus *Aspergillus awamori*: isolation and characterization of a feruloyl esterase and *p*-coumaroyl esterase. Enzyme Microb Technol 1994; 16:826-834.

McSweeney CS, Dulieu A, Webb RI, Dot TD, Blackall LL. Isolation and characterization of a Clostridium sp. with cinnamoyl esterase activity and unusual cell envelope ultrastructure. Arch Microbiol 1999; 172:139–149

Meinicke P. UFO: a web server for ultra-fast functional profiling of whole genome protein sequences. BMC Genomics 2009; 10:409

Moore J, Bamforth CW, Kroon PA, Bartolome B, Williamson G. Ferulic acid esterase catalyses the solubilization of β-glucans and pentosans from the starchy endosperm cell walls of barley. Biotechnology Letters 1996; 18:1423-1426

Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. Journal of Computational Chemistry 1998; 19(14):1639-1662

Moukouli M, Topakas E, Christakopoulos P. Cloning, characterization and functional expression of an alkalitolerant type C feruloyl esterase from Fusarium oxysporum. Appl Microbiol Biotechnol 2008; 79:245-254

Mukherjee G, Singh RK, Mitra A, Sen SK. Ferulic acid esterase production by *Streptomyces* sp. Biores Technol 2007; 98:211-213

Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1997; 10:1-6

Noble WS. What is a support vector machine?. Nat Biotechnol 2006; 24:1565-1567

Nsereko VL, Smiley BK, Rutherford WM, Spielbauer A, Forrester KJ, Hettinger GH, Harman EK, Harman BR. Influence of inoculating forage with lactic acid bacterial strains that produce ferulate esterase on ensilage and ruminal degradation of fiber. Anim Feed Sci Technol 2008; 145:122-135

Oili K, Markku HK. Distance Measures in the Training Phase of Self-Organizing Map for Color Histogram Generation in Spectral Image Retrieval. J Imag Sci Tech 2008; 52(2):020201-020500

Ong SA, Lin HH, Chen YZ, Li ZR, Cao Z. Efficacy of different protein descriptors in predicting protein functional families. BMC Bioinformatics 2007; 8:300

Pérez-Bercoff A, Koch J, Bürglin TR. LogoBar: bar graph visualization of protein logos with gaps. Bioinformatics 2006; 22(1):112-4.

Platt J. Machines using sequential minimal optimization. In Schoelkopf, B., Burges, C. & Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press, 1998

Prates JA, Tarbouriech N, Charnock SJ, Fontes CM, Ferreira LM, Davies GJ. The structure of the feruloyl esterase module of xylanase 10B from Clostridium thermocellum provides insights into substrate recognition. Structure 2001; 9(12):1183-90

Puchart V, Vršanská M, Mastihubová M, Topakas E, Vafiadi C, Faulds CB, Tenkanen M, Christakopoulos P, Biely P. Substrate and positional specificity of feruloyl esterases for monoferuloylated and monoacetylated 4-nitrophenyl glycoside. J Biotechnol 2007; 127:235–243

Ralet MC, CB Faulds, Williamson G, Thibault JF. Degradation of feruloylated oligosaccharides from suger-beet pulp and wheat bran by ferulic acid esterase from *Aspergillus niger*. Carbohydr Res 1994; 263:257-269

Rashamuse K, Burton S, Cowan D. A novel recombinant ethyl ferulate esterase from *Burkholderia multivorans*. J Appl Microbiol 2007: 103:1610–1620

Record E, Asther M, Sigoillot C, Pagès S, Punt PJ, et al. Overproduction of the *Aspergillus niger* feruloyl esterase for pulp bleaching application. Appl Microbiol Biotechnol 2003; 62:349-355

Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999; 12(2):85-94

Rumbold K, Biely P, Mastihubova M, Gudelj M, Gübitz G, et al. Purification and Properties of a Feruloyl Esterase Involved in Lignocellulose Degradation by *Aureobasidium pullulans.* Appl Environ Microbiol 2003; 69(9): 5622–5626

Sakamoto T, Nishimura S, Kato T, Sunagawa Y, et al. Efficient extraction of ferulic acid from sugar beet pulp using the supernatant of Penicillium chrysogenum. J Appl Glycosci 2005; 52:115-120

Sancho AI,Faulds CB, Bartolome B, Williamson G. Characterisation of feruloyl esterase activity in barley. J Sci Food Agric 1999; 79:447-449

Sankararaman S, Sjölander K. INTREPID-INformation-theoretic TREe traversal for Protein functional site IDentification. Bioinformatics 2008; 24(21):2445-2452

Sankararaman S, Kolaczkowski B, Sjölander K. INTREPID: a web server for prediction of functionally important residues by evolutionary analysis. Nucleic Acids Res 2009; 37:W390-395

Schneider G, Wrede P. The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. Biophys J 1994; 66:355-344

Schrödinger Suite 2009, Protein Preparation Wizard; Schrödinger, LLC, New York, NY, 2009.

Schubot FD, Kataeva IA, Blum DL, Shah AK, Ljungdahl LG, Rose JP, Wang BC. Structural basis for the substrate specificity of the feruloyl esterase domain of the cellulosomal xylanase Z from Clostridium thermocellum. Biochemistry 2001; 40(42):12524-12532

Shin HD, Chen RR. Production and characterization of a type B feruloyl esterase from *Fusarium proliferatum* NRRL 26517. Enzym Microb Technol 2006; 38:478-485

Shin HD, McClendon S, Le T, Taylor F, Chen RR. A Complete Enzymatic Recovery of Ferulic Acid From Corn Residues With Extracellular Enzymes From Neosartorya spinosa NRRL185. Biotechnol Bioeng 2006; 95(6):1108-15

Slavin JL. Mechanisms for the impact of whole-grain foods on cancer risk. J Am Coll Nutr 2000; 19:300S–307S

Smith DC, Bhat KM, Wood TM. Xylan-hydrolysing enzymes from thermophilic and mesophilic fungi. World J Microb Biot 1991; 7:475-484

Tarbouriech N, Prates JA, Fontes CM, Davies GJ. Molecular determinants of substrate specificity in the feruloyl esterase module of xylanase 10B from Clostridium thermocellum. Acta Crystallogr D Biol Crystallogr 2005; 61(2):194-197

Tenkanen M, Schuseil J, Puls J, Poutanen K. Production, purification and characterization of an esterase liberating phenolic acids from lignocellulosics. J Biotechnol 1991; 1:69-84

Teodoro ML, Phillips GN. Molecular docking: A problem with thousands of degrees of freedom, IEEE International Conference on Robotics and Automation, Seoul, Korea, IEEE 2001; 960-966

Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994; 22(22):4673-4680

Tomoko S, Koichi K, Tadashi I. Feruloyl esterases from suspension-cultured rice cells. Bulletin of FFPRI 2002; 385(4):225-23

Topakas E, Stamatis H, Mastihubova M, Biely P, Kekos D, Macris BJ, Christakopoulos P. Purification and characterization of a *Fusarium oxysporum* feruloyl esterase (FoFAE-I) catalysing transesterification of phenolic acid esters. Enzym Microb Technol 2003; 33:729–737

Topakas E, Stamatis H, Biely P, Kekos D, Macris BJ, Christakopoulos P. Purification and characterization of a feruloyl esterase from Fusarium oxysporum catalyzing esterification of phenolic acids in ternary water-organic solvent mixtures. J Biotechnol 2003; 102:33-44

Topakas E, Kalogeris E, Kekos E, Macris BJ, Christakopoulos P. Production and partial characterisation of feruloyl esterase by Sporotrichum thermophile in solid-state fermentation. Process Biochem 2003; 38: 1539- 1543

Topakas E, Stamatis H, Biely P, Christakopoulos P. Purification and characterization of a type B feruloyl esterase (StFAE-A) from the thermophilic fungus Sporotrichum Thermophile. Appl Microbiol Biotechnol 2004; 63:686–690

Topakas E, Christakopoulos P, Faulds CB. Comparison of mesophilic and thermophilic feruloyl esterases: Characterization of their substrate specificity for methyl Phenylalkanoates. J Biotechnol 2005; 115:355–366

Topakas E, Vafiadi C, Stamatis H, Christakopoulos P. *Sporotrichum thermophile* type C feruloyl esterase (StFaeC): purification, characterization, and its use for phenolic acid (sugar) ester synthesis. Enzym Microb Technol 2005; 36:729-736

Topakas E, Vafiadi C, Christakopoulos P. Microbial production, characterization and applications of feruloyl esterases. Process Biochem 2007; 42:497–509

Trzcińska M, Halasińska AG, Wetoszka U, Sieliwanowicz B. Possibility of applying Feruloyl esterase from Aspergillus niger A.n.8 for degradation of a cell wall complex in selected cereals. Pol J Food Nutr Sci 2005; 14/55(2):171-176

Tsuchiyama T, Sakamoto T, Fujita T, Murata S, Kawasaki H. Esterification of ferulic acid with polyols using a ferulic acid esterase from Aspergillus niger. Biochim Biophys Acta 2006; 1760:1071–1079

Tsuchiyama M, Sakamoto T, Tanimori S, Murata S, Kawasaki H. Enzymatic Synthesis of Hydroxycinnamic Acid Glycerol Esters Using Type A Feruloyl Esterase from *Aspergillus niger*. Biosci Biotechnol Biochem 2007; 71(10):2606-2609

Uestuen B, Melssen WJ, Buydens LMC. Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel. Chemometrics and Intelligent Laboratory Systems 2006; 81:29-40

Vafiadi C, Topakas E, Wong KKY, Suckling ID, Christakopoulos P. Mapping the hydrolytic and synthetic selectivity of a type C feruloyl esterase (StFaeC) from Sporotrichum thermophile using alkyl ferulates. Tetrahedron: Asymmetry 2005: 16:373–379

Vafiadi C, Topakas E, Christakopoulos P, Faulds CB. The feruloyl esterase system of *Talaromyces stipitatus*: Determining the hydrolytic and synthetic specificity of TsFaeC. J Biotech 2006; 125:210-221

Vafiadi C, Topakas E, Bakx EJ, Schols HA. Christakopoulos P, Structural characterisation by ESI-MS of feruloylated arabino-oligosaccharides synthesised by chemoenzymatic esterification, Molecules 2007; 12:1367–1375

Vafiadi C, Topakas E, Alderwick LJ, Besra GS, Christakopoulos P. Chemoenzymatic synthesis of feruloyl D-arabinose as a potential anti-mycobacterial agent. Biotechnol Lett 2007; 29:1771–1774

Vafiadi C, Topakas E, Alissandratos A, Faulds CB, Christakopoulos P. Enzymatic synthesis of butyl hydroxycinnamates and their inhibitory effects on LDL-oxidation, J Biotechnol 2008; 133: 497–504

Vafiadi C, Topakas E, Nahmias VR, Faulds CB, Christakopoulos P.Feruloyl esterase-catalysed synthesis of glycerol sinapate using ionic liquids mixtures. J Biotechnol 2009; 139:124-129

Wang J, Delabie J, Aasheim HC, Smeland E, Myklebost O. Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. BMC Bioinformatics 2002; 3:36

Wang X, Geng X, Egashira Y, Sanada H. Purification and Characterization of a Feruloyl Esterase from the Intestinal Bacterium *Lactobacillus acidophilus.* Appl Environ Microbiol 2004; 70(4):2367–2372

Wang X, Geng X, Egashira Y, Sanada H. Release of Ferulic Acid from Wheat Bran by an Inducible Feruloyl Esterase from an Intestinal Bacterium Lactobacillus acidophilus. Food Sci Technol Res 2005; 11(3):241-247

Wilkinson TN, Speed TP, Tregear GW, Bathgate RA. Evolution of the relaxin-like peptide family. BMC Evol Biol 2005; 5:14